Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment

J. Charles Alderson

Continuum

Diagnosing Foreign Language Proficiency

This page intentionally left blank

Diagnosing Foreign Language Proficiency

The Interface between Learning and Assessment

J. Charles Alderson

Continuum
The Tower Building, 11 York Road, London SE1 7NX
15 East 26th Street, New York, NY 10010

First published 2005

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 0-8264-8503-0 (hardback)

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress.

Typeset by Aarontype Limited, Easton, Bristol
Printed and bound in Great Britain by

Contents

Acknowledgements

No book is produced by the author alone, but this book could never have been conceived, let alone produced, had it not been for an enormous number of people throughout Europe. Although addressing the general topic of the diagnosis of foreign language proficiency and diagnostic testing in particular, in this book I have drawn heavily on the DIALANG Project (www.dialang.org) for my inspiration, examples and conclusions. Without DIALANG this book would not be here. I am extremely grateful to all who sponsored, designed, worked in and contributed to the DIALANG Project, which started in 1996 and came to the end of its public funding in 2004.

First and foremost, I must acknowledge the financial sponsorship of the European Commission, Directorate General for Education and Culture, through its Socrates programme, LINGUA (Action D). Without that support the Project would never have got off the ground, and the moral support and constant advice of Sylvia Vlaeminck and Paul Holdsworth of the Language Policy Unit of that Directorate were essential to keep us on track. The advice of the Expert Monitoring Group, acting on behalf of the Socrates Committee, was also indispensable. From the Technical Assistance Bureau, Gillian McLaughlin was our mentor, friend and source of good advice and red wine, and she was rewarded by becoming Project Manager in 2003!

Phase One of the Project (1996–1999) was coordinated by Professor Kari Sajavaara of the University of Jyväskylä, Finland, and the Scientific Coordinator was Professor Sauli Takala. Phases Two and Three of the Project (1999–2003 and 2003–2004) were coordinated at the Free University of Berlin, Germany, by Professor Wolfgang Mackiewicz. I am very grateful for their encouragement of the research that led to this book and their willingness to assist in any way possible.

Test Development within the DIALANG Project was very ably coordinated by Ari Huhta of the University of Jyväskylä, Finland, and José Noijons of CITO, the Netherlands.

Felianka Kaftandjieva was a highly competent adviser on statistical matters in Phase 1 and leader of the Data Analysis working group, which included Sauli Takala (Jyväskylä), Norman Verhelst and John de Jong (CITO). Timo Törmäkangas and Perttu Venermo provided helpful assistance with the data and computer systems. Data analysis was coordinated in Phase 2 by Caroline Clapham (Lancaster), Sari Luoma (Jyväskylä) and Norman Verhelst (CITO). A myriad of analyses as conducted by Tuija Hirvelä, who never complained of all the demands I made on her time for the supply of yet more data.

The self-assessment and feedback components were developed by a team of experts, coordinated by Alex Teasdale of Thames Valley University, who is sadly no longer with us, and aided by Neus Figueras of the Catalonian Department of Education, Sauli Takala, Felianka Kaftandjieva and Ari Huhta, all from Jyväskylä, Mats Oscarson of Göteborg University, Sweden and Steve Fligelstone from Lancaster. Brian North of Eurocentres, also representing the Council of Europe, acted as consultant to the group.

The IT team was based at the University of Lancaster, and included Steve Alexander, Steve Cotton, Randa El-Marakby, Tom Clapham, Jon Mathews, Adrian Fish and Jerry Treweek. Steve Childs was an excellent consultant on database matters. At Jyväskylä, Peppi Taalas, Perttu Venermo, and Jarkko Hänninen provided invaluable services. The IT team was initially managed by Steve Fligelstone, who later took on the arduous role of Development Coordinator. Graeme Hughes was a consultant to the IT team, and offered advice on numerous matters of system and program development, and was heavily involved in data management.

DIALANG was a large and complex Project, and the efforts of many, many individuals were crucial to success, although their individual hard work is rarely acknowledged. I am very happy to be able to record here all those who contributed to the Project: if I have forgotten anybody, I beg their forgiveness.

The Assessment Development Teams (ADTs) for the 14 languages, and the many translators, reviewers and standard-setting judges performed sterling work. In order of language, they were as follows:

Danish
John E. Andersen (ADT Leader), Anne Holmen, Esthi Kunz, Marie Linnet (Language Representative, Phase 2), Boris Sondersted, Annette Hagel Sorensen.

Dutch
Ron Oostdam (ADT Leader), Lut Baten, H. van Berlo, B. Bossers, R. Fukkink, M. Jetten, Henk Kuijper (Language Representative, Phase 2), P. Lukassen, José Noijons, E. van Schooten, H. Stortelder.

English
Richard West (ADT Leader), Donald Adamson, Jane Andrews, Patricia Aresnik, Jayanti Banerjee, Charles Boyle, Terry Bray, Deborah Cash, Caroline Clapham, Mark Crossey, S. Davies, J. Deachey, F. Fay, Gavin Floater, Alistair Fortune, M. Fulcher, Ari Huhta, R. Johnson, Lisa Lantela, Andrew Lewis, Sari Luoma, Julie Mezera, S. Panting, Anne Pitkänen-Huhta, Pirjo Pollari, Anne Räsänen, S. Shenton, Anne Sjöberg, Henna Tossavainen, Eleanor Underwood, Dianne Wall, Alan Walton.

Finnish
Mirja Tarnanen (ADT Leader), Eija Aalto, Marja Ahola, Pirjo Eskelinen-Leppänen, Ari Huhta (Language Representative, Phase 2), Hanna Immonen, Vesa Jarva, Eija Julkunen, Sirpa Kivelä, Tuija Lehtonen, Minna-Riitta Luukka, Ari Maijanen, Maisa Martin, Riitta Mielonen, Hilkka Miettinen, Johanna Peltola, Reija Portti, Kirsi Seppänen, Marja-Terttu Storhammar, Sirpa Tereska, Kirsti Tervo, Minna Torvinen-Artimo, Henna Tossavainen.

French
Chantal Rousseau (ADT Leader), Cécile Baumard, Catherine Benoît, Jérôme Bigo, Marie-José Billet, Yves Chevalier, Françoise Deparis, Caroline Duchateau, Martine Eisenbeis, Thomas Fraser, Abdi Kazeroni, Annette Locway, Dominique Perrin, Isabelle Seidel-Nicolas (Language Representative, Phase 2), Daniel Toudic. Actors for the audio-recordings were provided by the companies 'Guichet Unique' and 'Mahira'.

German
Johann Fischer (ADT Leader), Carl Braun, Markus Breyer, L. Bunn, Kathrin Burgel, Claudia Burghoff, Andrea Dötterer-Händle, Anke Ehlert, Maila Eichhorn, Roland Fischer, Manuela Glaboniat, Gerhard von der Handt, Matthias Hülsmann, B. Krefting, Beate Lachenmayer, Annette Lang, Erkki Lautsila, Gabriela Leder (Language Representative, Phase 2), Michael Lindenberg, Beate Maleska, Martina März, Nicola Reimann, Martina Rösch, Anja Scharnweber, Chirstel Schubert, Beate Schulze, Ulla-Britt Stiernskog-Migliore, Manfred Waitzbauer, W. Welter, Michael Werner.

Greek
Niovi Antonopoulou (ADT Leader and Language Representative, Phase 2), Panagiotis Azvaultis, George Bleris, Marianna Christou, Theodora Kaldi-Koulikidou, Theresa Madagan, Demetro Manavi, Constantinos Matis, Maria Moumtzi, Andreas Papapavlou, Pavlos Pavlou, Heleni Prodromidou, Vasso Tocatlidou, Antonios Tsopanoplou, Maria Tsoureli, Zapheroulo Vassiliou, Smaro Voyiatzidou.

Icelandic
Svavar Sigmundsson (ADT Team Leader), Gerda Cook, María Garðarsdóttir, Dóra Hjartardóttir, Thora Magnúsdóttir, Kai Saanila, Eyjólfur Sigurdsson (Language Representative, Phase 2). Sigríður Þorvaldsdóttir, Peter Weiss (Language Representative, Phase 2).

Irish
Tadhg Ó hIfearnáin (ADT Leader), David Barnwell, Annette Byrd, Angela Chambers, Jean Conacher, Diarmuid Ó Gruagáin, Muiris Ó Laoire, Eoghan Mac Aogáin (Language Representative, Phase 2), Anthony McCann, Siobhan Murphy, Nóilin Nic Bhloscaidh, Angela Rickard.

Italian
Gabriella Pavan de Gregorio (ADT Leader and Language Representative, Phase 2), Luigia Acciaroli, W. D'Addio, L. De Bellis, Raimondo Bolletta, Franco Capparucci, Suzanne Ely, Carola Feltrinelli, Fernando Filoni, Francesco Fulminanti, P. Gensini, S. Gensini, P. Giunchi, P. Giunti, Enrico Grazzi, Rosaria Lustrissimi, Rita Marconi, P. Mezzaluna, Carlo Odorico, Nadia Persiani, Giorgio Possamai, Graziella Pozzo, Simonetta Rossi, Claudia Scotese, Silvana Serra, R. Titone, Sebastiano Triulzi, Paola Tulliani, Rita Valente, Andrea Villarini.

Norwegian
Reidun Oanes Andersen (ADT Leader and Language Representative, Phase 2), Karina Andersen, Cecilie Carlsen, Erna Crowo, Sigrun Eilertsen, Gölin Evertsen, Anne Golden, Jon Erik Hagen, Angelvik Halvard, Berit Halvorsen, Arvid Hellum, Grethe Hilditch, Tore Hoyte, Berit Lie, Kirsti MacDonald, Turid Mangerud, Jorn Pedersen, Nils Ree, Lisbeth Salomonsen, Svavar Sigmundson.

Portuguese
Ana Cristina M. Lopes (ADT Leader), Filomena Marques de Carvalho, Edite Ferreira, Lucilia Ferreira, Sofia Claudia Gomes, Paulo Melo, Maria José Moura Santos, Isabel Pereira (Language Representative, Phase 2), Graca Rio-Torto, Antonino Silva, Diana Silver, Vitor Torres, Graca Trindade.

Spanish
Fuesanta Puig (ADT Leader), Arlette Aubier, Rocio Barros, M. Carmen Ferris, Neus Figueras (Language Representative, Phase 2), Angels Oliveras, Lourdes Perdigo, Consol Perez, Juan A. Redo, Rosa Rialp, Pilar Utrero.

Swedish
Barbro von Elek (ADT Leader), Annika Bergström, Ulric Björck, Marianne Demaret, Tibor von Elek, Anita Forsmalm, Stefan Nordblom,

Mats Oscarson (Language Representative, Phase 2), Rauni Paakkunainen, Leena Skur, Sylvi Vigmo.

The partner institutions represented in the ADTs for the 14 DIALANG languages were the following:

Language	*Institution*
Danish:	Københavns Universitet
Dutch:	Universiteit van Amsterdam
English:	Victoria University of Manchester
Finnish:	Jyväskylän yliopisto
French:	Université Charles de Gaulle – Lille 3
German:	Universität Hohenheim
Greek:	Aristoteleio Panepisimio Thessalonikis
Icelandic:	Háskóli Ísland
Irish:	Ollscoil Luimnigh
Italian:	Centro Europeo dell'Educazione, CEDE
Norwegian:	Universitetet i Bergen
Portuguese:	Universidade de Coimbra
Spanish:	Universitat de Barcelona
Swedish:	Göteborgs Universitet

When in October 1998 the management structure of the project was changed, development work in this area was coordinated by the Test Development Working Group, consisting of José Noijons (CITO), team leader, Ari Huhta and Sari Luoma (both from Jyväskylä). In addition, Gabriella Pavan de Gregorio (CEDE) and Johann Fischer (University of Hohenheim) represented those partner institutions which were responsible for item production.

Paul Meara, University of Swansea, and his team, produced the Vocabulary Size Placement Tests in the 14 languages.

The Project Management Team for Phase 1 in Jyväskylä included Kari Sajavaara (Coordinator), Anu Halvari, Jarkko Hänninen, Eija Hietala, Tuija Hirvelä, Ari Huhta, Tarja Jukkala, Felianka Kaftandjieva, Liisa Kelloniemi, Marleena Kojo, Sari Komulainen, Sari Luoma, Maijaliisa Majamäki, Helen Niskanen, Soile Oikkonen, Anne Räsänen, Leena Skur, Peppi Taalas, Sauli Takala, Timo Törmäkangas, Perttu Venermo, Alan Walton.

The Project Management Team for Phases 2 and 3 in Berlin included Wolfgang Mackiewicz (Coordinator of Phase 2), Sabine Kinzius, Gillian McLaughlin, Ernst Miarka, Astrid Peter, Konrad Stransky, Alan Walton.

Dialang website development and maintenance was undertaken in Finland by the Mediakettu company: Jari Peurajärvi, Jouni Takkinen; and in Berlin by Ernst Miarka.

The following institutions made major contributions to the development of the DIALANG assessment system:

- Johannes Kepler Universität Linz – Austria
- Universität Wien – Austria
- Institut für Theoretische und Angewandte Translationswissenschaft, University of Graz – Austria
- Université Libre de Bruxelles – Belgium
- Handelshojskolen i Århus – Denmark
- Københavns Universitet, Institut for Nordisk Filologi – Denmark
- Université Charles de Gaulle, Lille 3 – France
- DIE, Frankfurt – Germany
- Universität Hohenheim, Stuttgart – Germany
- Aristoteleio Panepisimio Thessalonikis – Greece
- Háskoli Ísland, Reykjavik – Iceland
- Ollscoil Luimnigh, Limerick – Ireland
- Institttiúid Teangeolaíochta Éireann – Ireland
- CEDE, Frascati – Italy
- SCO-Kohnstamm Instituut; Universiteit van Amsterdam – Netherlands
- Universitetet i Bergen – Norway
- Universidade de Coimbra – Portugal
- Departamento de Línguas e Culturas, Universidade de Aveiro – Portugal
- Generalitat de Catalunya/Escoles Oficials d'Idiomes – Spain
- Göteborgs Universitet – Sweden
- Thames Valley University, London – UK
- Victoria University of Manchester – UK

The following institutions assisted the development of the DIALANG assessment system by participating in the trialling of the language tests:

- Vrije Universiteit Brussel – Belgium
- Århus Kommunes Sprogcenter – Denmark
- Århus Universitet – Denmark
- Helsingin yliopiston kielikeskus – Finland
- Jyväskylän ammatillinen aikuiskoulutuskeskus – Finland
- Kuopion aikuiskoulutuskeskus – Finland
- Edupoli, Porvoo – Finland
- Oulun yliopisto – Finland
- Pohjois-Savo Polytechnic, Kuopio – Finland
- Tampereen yliopiston kielikeskus – Finland
- Université de Technologie de Compiègne – France
- Rijksuniversiteit Groningen – Netherlands
- ESADE – Spain
- Universidad Autónoma de Madrid – Spain
- Lunds universitet – Sweden
- Språk- och kommunikationscenter, Utvecklingsavdelningen, Utbildningsförvaltningen, Stockholm – Sweden

- Université de Lausanne – Switzerland
- Centre de langues, Université de Lausanne – Switzerland

I was fortunate enough to be awarded a Leverhulme Research Fellowship in 2003 which enabled me to devote my time to sorting out the data, conducting analyses and writing reports, and beginning the writing of this book. I am very grateful to the Leverhulme Trust for their support, and to the Department of Linguistics and English Language, Lancaster University, which granted me a sabbatical term and tolerated my numerous absences during the final stages of this book. My colleagues Jay Banerjee and Dianne Wall took over my teaching and administrative responsibilities during my absences, for which I am deeply grateful, and I greatly appreciate their interest in, and support of, the Project, together with many members of the Language Testing Research Group at Lancaster.

I owe a very special debt of gratitude to Ari Huhta, of the University of Jyväskylä, and Graeme Hughes, of Lancaster University, for their constant willingness to supply data during the tenure of my Research Fellowship and beyond, to answer questions, to read drafts of chapters for their accuracy, and for being excellent colleagues and good friends.

I was especially fortunate to have a superb and patient Editor in Jenny Lovel, of Continuum Books, who provided speedy responses to queries, shepherded the book through the approval process with amazing alacrity and skill, and who showed great faith in me and the value of this book.

Last but not least, I am very grateful to Edit Nagy for her constant support, encouragement, advice and friendship.

J. Charles Alderson
Lancaster, December 2004

The publisher would like to thank Hodder Arnold for permission to reprint sections of 'The development of a suite of computer-based diagnostic tests based on the Common European Framework', J. Charles Alderson and Ari Huhta, *Language Testing*, 22(3), June 2005, Hodder Arnold.

Introduction

This book addresses an age-old topic, studies of which are still in their infancy, as the topic is under-researched and not well theorized or understood: that of the diagnosis of language proficiency. The word 'diagnosis' is common in discussions in language education and applied linguistics and the term 'diagnostic test' is not unusual in discussions of language testing. Such mentions, however, lack exemplification (much less explanation) and even definitions of the term fail to go beyond the superficial 'identification of strengths and weaknesses and their remediation'. As we shall see, even a rigorous field like foreign language testing lacks examples of diagnostic tests, and indeed is frequently confused as to whether it is even possible to devise tests which can diagnose aspects of foreign language proficiency. Related areas like second language acquisition research are still grappling with the fundamental questions of how foreign and second language proficiency develops, and therefore what underlies such development.

Yet at the heart of teaching and assessing foreign language proficiency lies the need to help learners make progress. If researchers, theoreticians and testers do not know how language proficiency develops, they can hardly claim to be able to help learners develop such an ability.

Clearly learners do develop in their proficiency, although few reach similar levels in a foreign language as they have attained in their native language, at least under conditions of instruction. How is this possible? Recent theories have suggested that the way learners develop is at least as much a function of exposure to the foreign language as it is a function of our theoretical or even practical understanding of what develops and how.

So we are faced with a paradox: learners do learn foreign languages, teachers and language educationalists believe they have some understanding of how learners might be encouraged to learn, and the phrase 'diagnosis in order to improve learning' is frequently used. But we lack descriptions in any detail of what changes as learners develop, there is little research underpinning what descriptions there are, and there is a

remarkable lack of valid diagnostic tests or indeed of any tests that claim explicitly to be diagnostic of foreign language proficiency.

It might therefore seem that this book is over-ambitious in its title and its content. The aim, however, is not to explain exactly how and when to diagnose a learner's foreign language proficiency, but rather to put the need for discussions of diagnosis firmly on the agenda of language testers, language educators, applied linguists and second language acquisition researchers. It aims to start a debate about how diagnostic testing might most appropriately be developed. I argue that the field has neglected to construct diagnostic tests, partly because other forms of testing – in particular high-stakes testing – have dominated the field. I am not alone in making this claim, but I also argue that even those who would concentrate their efforts on understanding classroom assessment procedures have failed to address the need for diagnosis of learners' strengths and weaknesses.

This book is not, however, simply a discussion of the need for better diagnosis. Rather, it is based on attempts to develop computer-based diagnostic tests, through a European Union-funded project known as DIALANG. I concentrate on describing and exemplifying the diagnostic system that has been developed in this project, and on exploring the empirical results to date. DIALANG represents a unique opportunity to begin to come to grips with the notion of diagnosis, to explore the limits of what is known to date, and to develop an agenda for further research into the possibilities of the DIALANG system for helping us understand better the nature of foreign language proficiency, and for developing better, more focused and more theoretically informed assessment procedures that might eventually lead to improved diagnosis of the development of foreign language proficiency.

There are those who claim that they already know how foreign language proficiency develops, in particular aspects of the grammatical or morphological systems of some languages, just as there are others who claim that they have already developed scales of language proficiency that show how learners progress. But both sets of claims are inadequate: the former, second language acquisition researchers, because they only focus on narrow aspects of such a complex topic as foreign language proficiency, involving micro-studies which have nothing to say about macro-issues like proficiency development; and the latter, proficiency testers, because they fail to show how real learners do indeed develop, and thus the claims of their scales lack validation.

Nevertheless, we will have made progress in our understanding of diagnosis and diagnostic testing if the two fields of second language acquisition and language testing begin to engage in debate about the need for and the nature of diagnosis, from both micro- and macro-points of view. The history of collaboration between the two fields is not encouraging: although repeated calls have been made for such collaboration (see Bachman and Cohen, 1998 or Banerjee and Franceschina,

2004), there has been little activity to date. But it is my hope that the discussions in this book will spark debate, controversy perhaps, but above all attempts to understand better what needs to be diagnosed and what can be diagnosed, and may lead to attempts to develop better diagnostic instruments than the ones that currently exist.

Thus it is my hope also that this book might assist the age-old but infant field of diagnostic foreign language testing to take a few faltering steps on the long road to maturity. I believe that DIALANG has risen to the challenge, and has shown us how we might move forward, as well as how much remains to be understood. I believe that further research using the DIALANG system will contribute to that understanding. DIALANG is unique in that it attempts the diagnostic assessment of 14 European languages, not merely one dominant language. Although this book focuses almost exclusively on English, that is simply because there is more data available to date on the performance of the English tests in DIALANG. Yet as DIALANG is freely available over the Internet, it is in principle possible to use it as a tool for the understanding of how proficiency in other languages develops, including Icelandic and Irish, Greek and German, Spanish and Swedish, and hence what can be learned about the diagnosis of proficiency in those languages too.

Of course, computer-based diagnosis is limited in what it can achieve: it can tell us little or nothing about how an ability to speak a foreign language develops. But a limited start to our understanding of such a complex topic is better than no start at all, and it is my contention that it is about time we started to understand.

Chapter 1: Diagnosing foreign language proficiency: a teaching/testing interface

Introduction

It is a commonplace to claim the importance of assessment in language teaching and learning. Teachers need to know what learners already know, what they have learned in the course of instruction over a longer or shorter period and where their strengths and weaknesses are, so that they can plan their instruction appropriately, guide learners on where they need to improve and give feedback to learners. Unfortunately, when teachers give students tests, it usually takes a few days, if not longer, for learners to receive their results, and so the feedback lacks immediate relevance. Moreover, tests made by teachers are often of poor quality, and the insight they could offer into achievement, progress, strengths and weaknesses is usually very limited indeed. In the case of national and international examinations and proficiency tests, feedback may be delayed by several months, and the results are irrelevant to learning needs, with little or no information to help learners understand what they need to do in order to improve.

The type of test that comes closest to being central to learning is the diagnostic test.

What is a diagnostic test?

Language testing handbooks frequently distinguish diagnostic tests from placement, progress, achievement and proficiency tests. The ALTE multilingual glossary defines a diagnostic test thus: '*A test which is used for the purpose of discovering a learner's specific strengths or weaknesses. The results may be used in making decisions on future training, learning or teaching*' (ALTE, 1998). However, diagnostic tests are frequently confused with placement tests. The same multilingual glossary defines 'placement test' as follows: '*A test administered in order to place students in a group or class at a level appropriate to their degree of knowledge and ability*' (*op.cit.*). According to this,

both diagnostic and placement tests appear to be designed to identify what a learner knows in order to decide on future teaching or learning. Placement tests, after all, are intended to group learners in homogeneous groups in order to have a suitable basis for further teaching and learning.

The Davies *et al.* dictionary of language testing describes the use of diagnostic tests as follows:

> *information obtained from such (diagnostic) tests is useful at the beginning of a language course, for example, for placement purposes (assigning students to appropriate classes), for selection (deciding which students to admit to a particular course), for planning of courses of instruction or for identifying areas where remedial instruction is necessary.* (Davies *et al.*, 1999)

In other words, diagnostic tests are used for placement purposes – and thus appear to be identical to placement tests. Indeed, Davies *et al.* continue: '*It is common for educational institutions (e.g., universities) to administer diagnostic language tests to incoming students, in order to establish whether or not they need or would benefit from support in the language of instruction used*' (*op.cit.*). In short, there appears to be no distinction between a diagnostic test and a placement test, at least according to these authors.

It is arguable whether it is indeed the case that 'it is common for universities to administer diagnostic language tests'. Much more common is the administration of some form of post-admissions placement test (for an account of such a placement test and the problems of validating it, see Wall *et al.*, 1996). I have never heard any university claim that what it is doing is diagnosing students' strengths and weaknesses with its placement procedures, since typically what is decided is who needs some form of in-sessional or pre-sessional assistance, and who does not. At most, what is decided is whether a student should take a reading class or a grammar class, a writing class or a listening class or, in the case of US institutions, whether a student should take only a fractional load of academic courses rather than a full academic load. This is rarely known as diagnosis.

Bachman (1990) asserts that

> *virtually any language test has some potential for providing diagnostic information . . . Information from language tests can be used for diagnosing students' areas of strength and weakness in order to determine appropriate types and levels of teaching and learning activities . . . A placement test can be regarded as a broad-band diagnostic test in that it distinguishes relatively weak from relatively strong students so that they can be provided learning activities at the appropriate level. Similarly, a readiness test differentiates students who are ready for instruction from those who are not. A detailed analysis of student responses to the questions on placement and readiness tests can also provide more specific information about particular areas of weakness.* (p. 60)

Alderson *et al.* (1995) define diagnostic tests similarly:

> *Diagnostic tests seek to identify those areas in which a student needs further help. These tests can be fairly general, and show, for example, whether a student needs particular help with one of the four main language skills; or they can be more specific, seeking perhaps to identify weaknesses in a student's use of grammar. These more specific diagnostic tests are not easy to design since it is difficult to diagnose precisely strengths and weaknesses in the complexities of language ability. For this reason there are very few purely diagnostic tests. However, achievement and proficiency tests are themselves frequently used, albeit unsystematically, for diagnostic purposes.* (p. 12)

Thus it would appear that even achievement and proficiency tests can perform diagnostic functions. Indeed, Davies *et al.* claim that '*relatively few tests are designed specifically for diagnostic purposes*'. They explain this as follows: '*It is difficult and time-consuming to construct a test which provides detailed diagnostic information*'.Yet it is somewhat odd to say that few diagnostic tests exist because they are time-consuming to construct – proficiency tests are also difficult and time-consuming to construct, but many exist.

It would appear that we have a problem here: diagnosis is useful, most language tests can be used for diagnosis in some sense, it is common for universities to administer diagnostic tests, and yet diagnostic tests are rare!

To summarize: there is considerable confusion in the literature between placement tests and diagnostic tests. Furthermore, it is frequently claimed that achievement and proficiency tests can be used for diagnostic purposes. Yet it is said that diagnostic tests are rare, and are very difficult to construct. Moreover, there are frequent references in the language testing research literature to, and investigations of, proficiency tests, achievement tests, placement tests and even aptitude tests but diagnostic tests are very rarely referred to or investigated.

The reason for this is not immediately apparent. Clearly it would be useful to have tests which enable teachers and learners to diagnose their strengths and weaknesses. However, linguistic diagnosis might not be as well developed as is desirable because of the lack of attention paid to the subject. High-stakes tests that directly affect people's lives, like university entrance tests or proficiency tests for citizenship or employment, have more obvious potentially negative consequences, and so the quality of such instruments is of paramount importance. Inadequate diagnosis in the context of language education is unlikely to be life-threatening, unlike inadequate medical diagnosis. And so much less attention has been devoted to ensuring the validity and reliability of diagnostic tests in the foreign language field.

The aim of this book is to describe and discuss the nature of diagnostic tests, in order to understand better how they might be constructed and validated.

What might diagnostic tests contain?

Bachman (1990) offers the following thoughts on what is usually considered to be suitable content for diagnostic tests:

> *When we speak of a diagnostic test . . . we are generally referring to a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency. Thus, diagnostic tests may be either theory or syllabus-based.* (p. 60)

According to Bachman, a diagnostic test might contain important aspects of the content of a specific programme or, alternatively, it might be based on a specific theory of language proficiency. However, the former is usually regarded as an achievement test, and the latter as a proficiency test. We appear to be little nearer specifying the content of diagnostic tests *per se*.

Moussavi (2002) has quite a long section on diagnostic testing, going into more detail about how achievement and proficiency tests can be used for diagnosis, with ideas on what might be suitable content:

> *If diagnostic testing is defined as providing feedback to teachers and students regarding their strengths and weaknesses, then almost any test would be diagnostic. Diagnostic tests are an essential part of individualized instruction programmes. In this case, the students take the test for each unit when they feel ready for them. The tests help them to see whether or not they are ready to move on to the next unit, assignment, or passage, and enable the teacher to develop a cumulative grade or credit. The diagnostic test tries to answer the question: How well have the students learnt this particular material? Since it relates to particular elements in the course which have just been taught, for example, 'type III conditional sentence with if' or 'asking permission', the assessment will give immediate feedback to the student. If his learning has been successful, the results will give a considerable lift to the student's morale and he is likely to approach the next learning tasks with fresh enthusiasm. The degree to which a test is diagnostic depends not so much on the purpose of the test, but on the way in which scores are analysed.*
>
> *Let us consider TOEFL for a moment: TOEFL is usually considered a PROFICIENCY TEST, and when its total score is considered by an admissions officer, it can quite rightly be so classified. However, if one looks at the five part-scores for reading comprehension, vocabulary, and so on, the test is serving a diagnostic purpose in that information about an individual's particular strengths and/or weaknesses is obtained. That is, we have specific information not on 'English', but on certain abilities or skills. As information from these tests can be used for diagnosing students' areas of strength and weakness in order to determine appropriate types and levels of teaching and learning activities, therefore virtually any language test has some potential for providing diagnostic information. Diagnostic tests are the reverse side of ACHIEVEMENT TESTS in the sense that*

> *while the interest of the achievement test is in success, the interest in the diagnostic test is in failure, what has gone wrong, in order to develop remedies.*

Like Bachman, however, Moussavi has not clearly identified the difference in content between diagnostic tests on the one hand, and achievement and proficiency tests on the other hand.

Bachman and Palmer (1996) have the following to add to this discussion of diagnosis:

> *Diagnosis involves identifying specific areas of strength or weakness in language ability so as to assign students to specific courses or learning activities. For example, if a language program included three different courses, one focused on the editing of sentence level grammar and punctuation errors, a second focused on revising the organisation of essays, and a third focused on logic of argumentation in writing, a teacher might use a test that included all these different language use activities as a basis for deciding which course would be most appropriate for students to take.* (p. 98)

Thus, a diagnostic test would appear to be much more specific, focused and related to particular (and different) language programmes. In contrast, their description of a placement test merely talks about assigning students to different 'levels' of a language course.

This apparent distinction is not, however, sustained throughout the Bachman and Palmer volume. Later, they present an extended description of the development of a 'syllabus-based diagnostic achievement test' for students in an ESP programme. In fact, this 'partially developed example' is rather similar to other examples in the book, the main difference being that '*we will now control more tightly the exact structures that we want the test takers to use*'. Curiously, however, the specific purpose of the test is to determine whether students have mastered specific course content, and will provide diagnostic feedback to those who have not, as well as to course designers and teachers, who can then tailor the course more closely to the needs of the students. It is not clear how this is different from an achievement test.

Detailed content specifications are given, focusing on prepositions, participial modifiers, personal and relative pronouns, verb forms, infinitives, adverbs, quantifiers, auxiliary verb forms and coordinating and subordinating conjunctions, as well as punctuation marks like comma, semicolon, period and question mark. Authenticity is acknowledged to be limited (might this be a feature of diagnostic tests?), the test method is limited and it is suggested that '*there are no obvious affective barriers*' that would prevent students performing at their best. This might indeed be a feature of diagnostic tests, since they are sometimes said to be low-stakes. Unfortunately, although feedback to students, teachers and course designers is said to be important, no indication is given as to how the test results will be reported, However, for a diagnostic test

to be truly useful, profiles of performance are surely needed, and very detailed information on the performance across the various components specified in the content specifications is highly desirable.

Hughes (1989) considers that the diagnosis of linguistic strengths and weaknesses is very difficult:

> *It is not so easy to obtain a detailed analysis of a student's command of grammatical structures, something which would tell us, for example, whether she or he had mastered the present perfect/past tense distinction in English. In order to be sure of this, we would need a number of examples of the choice the student made between the two structures in every different context which we thought was significantly different and important enough to warrant obtaining information on. A single example of each would not be enough, since a student might give the correct response by chance. As a result, a comprehensive diagnostic test of English grammar would be vast (think of what would be involved in testing the modal verbs, for instance). The size of such a test would make it impractical to administer in a routine fashion. For this reason, very few tests are constructed for purely diagnostic purposes, and those that there are do not provide very detailed information.*
>
> *The lack of good diagnostic tests is unfortunate. They could be extremely useful for individualised instruction or self-instruction. Learners would be shown where gaps exist in their command of the language, and could be directed to sources of information, exemplification and practice. Happily, the ready availability of relatively inexpensive computers with very large memories may change the situation. Well-written computer programs would ensure that the learner spent no more time than was absolutely necessary to obtain the desired information, and without the need for a test administrator. Tests of this kind will still need a tremendous amount of work to produce. Whether or not they become generally available will depend on the willingness of individuals to write them and of publishers to distribute them.* (pp. 13–14)

Whilst repeating what Davies *et al.* say about placement, Hughes goes further to discuss why it might be difficult to design tests which could provide detailed diagnoses, and to suggest that computer-based testing might offer a solution. (Interestingly, Hughes predicts the advent of computer-based diagnostic testing, eight years before the DIALANG Project began.) He does not, however, address the issue of the construct: what exactly should a diagnostic test of grammar, say, actually contain? And one of reading, or speaking?

Interestingly, in the second edition (Hughes, 2003) there are more, albeit somewhat brief, references to diagnostic tests as follows:

> *Diagnostic tests of grammar . . . will tend to be discrete-point.* (p. 19)

> *It may also be thought worthwhile testing lower level listening skills in a diagnostic test, since problems with these tend to persist longer than they do in reading. These might include:*

- *discriminate between vowel phonemes*
- *discriminate between consonant phonemes*
- *interpret intonation patterns (recognition of sarcasm, questions in declarative form, etc., interpretation of sentence stress).* (p. 162)

The usefulness (and indeed the feasibility) of a general diagnostic test of vocabulary is not readily apparent. As far as placement tests are concerned, we would not normally require, or expect, a particular set of lexical items to be a prerequisite for a particular language class. All we would be looking for is some general indication of the adequacy of the students' vocabulary. The learning of specific lexical items in class will rarely depend on previous knowledge of other, specified items. (p. 179)

It is encouraging for the future of diagnostic testing that Hughes has begun to address the topic in more depth than other authors. However, his views on the relative ease of diagnosing language use skills and the difficulty of diagnosing grammatical and lexical knowledge remain mere suppositions. It is far from clear why it should be possible to diagnose strengths and weaknesses in reading or listening, for example. The warning about the need to test grammatical structures in a comprehensive range of contexts must surely, *a priori*, also apply to a need to ascertain a learner's ability to process a whole range of different texts across a wide variety of topics at varying levels of linguistic difficulty. Conversely, until we have developed tests of grammatical knowledge across the range of contexts which Hughes asserts are essential, we will not know whether it is indeed essential to include so much variation, or whether many of the items/contexts are redundant. However speculative and uninformed by empirical evidence though these comments may be, they are helpful in constituting a set of hypotheses which can be tested empirically. Thus can diagnostic tests be explored, and our understanding enhanced.

In summary, the language testing literature offers very little guidance on how diagnosis might appropriately be conducted, what content diagnostic tests might have, what theoretical basis they might rest on, and how their use might be validated. Diagnostic testing is virtually ignored in the literature, and is rarely, if ever, problematized. Nevertheless, we have seen some sporadic indications in the literature that diagnostic tests might test aspects of the control of the linguistic elements, especially for morphosyntax and phonemes. What might be tested in detail in the diagnosis of strengths and weaknesses in language use skills is much less clear.

Towards a characterization of diagnostic tests

Despite the lack of detailed discussion in the language testing literature, I suggest that it might indeed be possible to begin to identify some

features of diagnosis that might distinguish diagnostic tests from other types of test, even if other tests continue to be used for diagnostic purposes – or pseudo-diagnostic purposes like placement. I list below a set of hypothetical features of diagnostic tests, as a synthesis of the literature reviewed above, which might guide further thinking about, design of and research into diagnostic tests. It should be pointed out, however, that some of these 'features' contradict other features, and, more importantly, that they constitute a potential agenda for research rather than a set of definitive statements about what is necessary and possible.

1 Diagnostic tests are designed to identify strengths and weaknesses in a learner's knowledge and use of language.
2 Diagnostic tests are more likely to focus on weaknesses than on strengths.
3 Diagnostic tests should lead to remediation in further instruction.
4 Diagnostic tests should enable a detailed analysis and report of responses to items or tasks.
5 Diagnostic tests thus give detailed feedback which can be acted upon.
6 Diagnostic tests provide immediate results, or results as little delayed as possible after test-taking.
7 Diagnostic tests are typically low-stakes or no-stakes.
8 Because diagnostic tests are not high-stakes they can be expected to involve little anxiety or other affective barriers to optimum performance.
9 Diagnostic tests are based on content which has been covered in instruction, or which will be covered shortly.

OR

10 Diagnostic tests are based on some theory of language development, preferably a detailed theory rather than a global theory.
11 Thus diagnostic tests need to be informed by SLA research, or more broadly by applied linguistic theory as well as research.
12 Diagnostic tests are likely to be less 'authentic' than proficiency or other tests.
13 Diagnostic tests are more likely to be discrete-point than integrative, or more focused on specific elements than on global abilities.
14 Diagnostic tests are more likely to focus on language than on language skills.
15 Diagnostic tests are more likely to focus on 'low-level' language skills (like phoneme discrimination in listening tests) than higher-order skills which are more integrated.
16 Diagnostic tests of vocabulary knowledge and use are less likely to be useful than diagnostic tests of grammatical knowedge and the ability to use that knowledge in context.

17 Tests of detailed grammatical knowledge and use are difficult to construct because of the need to cover a range of contexts and to meet the demands of reliability.
18 Diagnostic tests of language use skills like speaking, listening, reading and writing are (said to be) easier to construct than tests of language knowledge and use. Therefore the results of such tests may be interpretable for remediation or instruction.
19 Diagnostic testing is likely to be enhanced by being computer-based.

To summarize and conclude this chapter, I have argued that the literature on diagnostic testing in second and foreign languages is inadequate, contradictory and confused. No clear distinction is made between diagnostic tests and placement tests: it is claimed by some that any test can be used for diagnostic purposes, and by others that it is very difficult to write diagnostic tests. It is argued that very few diagnostic tests of foreign language proficiency exist, and the reason for that may in part be because the very concepts of diagnosis and the nature of foreign language proficiency are undeveloped, unproblematized and undertheorized. The lack of diagnostic tests may also be due to more practical matters, namely that to date there has been no demand for, and little funding to support, the development of such tests.

However, with the advent of computer-based tests, especially those delivered over the Internet, it is now relatively straightforward to give learners immediate feedback on their performance, thereby making it possible for that feedback to have maximum impact, because learners may still recall their reasons for responding the way they did and be more receptive to the feedback and able to incorporate it into their developing interlanguage. Thus feedback becomes, potentially, maximally informative and relevant.

If tests are informed by an adequate theory of language use, language development and language learning, and if learners can receive feedback on their performance and their ability immediately, then the possibility for the incorporation of assessment into language learning becomes apparent, indeed urgent.

But first, it is important to look outside the field of second and foreign language testing to see whether diagnosis is as poorly developed a field in other content areas as it appears to be in the field of foreign language testing. In the next chapter, I shall examine the use of diagnostic tests in first language development and in learning to read.

Chapter 2: Diagnosis in other fields

The tradition in diagnosis

We saw in Chapter 1 that relatively little has been written about diagnostic testing in second and foreign language learning. Although most handbooks on language testing include diagnostic tests as a separate category of test, alongside proficiency, placement, achievement, progress and aptitude tests, the definitions of 'diagnostic test' frequently overlap with definitions of other types of test, and there is a degree of confusion in the literature about what exactly diagnostic tests are and how they should be constructed.

Indeed, perhaps the most notable thing about diagnostic testing within second or foreign language testing is the scarcity and brevity of its treatment, the lack of exemplification and advice on the construction of diagnostic tests, and the complete absence of any discussion of how diagnostic tests should or could be validated. In contrast, diagnostic tests of reading and learning difficulties in one's first language are reasonably well established, with quite a long history, and are widely used, at least in North America and Western Europe.

A review of literature in general education reveals that diagnostic testing has traditionally concentrated on the diagnosis of speech and language disorders (Nation and Aram, 1984), the detection of difficulties in reading and arithmetic (Clay, 1979; Bannatyne, 1971; Schonell and Schonell, 1960) and the diagnosis of learning difficulties (Bannatyne, 1971; Wilson, 1971). For example, Schonell and Schonell (1960) is entitled *Diagnostic and Attainment Testing*, and provides a guide to a range of diagnostic tests, of reading and spelling, and of arithmetic and 'English'.

Interestingly, there are few more recent books or indeed articles available on diagnostic testing outside the context of clinical speech and language pathology and diagnosis, and in foreign language testing specifically, as we have seen, there are no books on educational diagnostic language testing, very few articles, and the topic receives only cursory treatment even in textbooks on language testing. Yet most writers on

educational diagnosis and diagnostic testing emphasize the complexity of the task, the need to take account of a range of factors – background, motivation, memory, cognition, perception, not just test performance – that might impact on a student's learning or abilities, and the crucial importance of the diagnostician or the diagnostic test constructor having a precise, thorough and clear understanding of the nature of 'normal' behaviours and abilities as well as the processes underlying those behaviours and the manifestations of the abilities.

Bannatyne remarks, in the context of the diagnosis of language, reading and learning difficulties: '*We owe these children the courtesy of a long (though in short sessions), extensive and thorough examination which should be no less efficient than the equivalent diagnostic examination they would receive in a large hospital for an obscure physical disease. As professionals, we all have to progress beyond the "IQ assessment and reading test" approach, which is akin to the doctor taking a pulse and feeling the brow*' (1971: 630).

Schonell and Schonell (1960) say that '*the diagnostic test is constructed not to assess levels but to reveal difficulties in school subjects*' (admitting in a footnote that errors from attainment tests may provide useful diagnostic information). They go on to say that the diagnostic test '*is based upon an analysis of the mental processes involved in learning a subject – as in learning to read or to spell – or an analysis of the material required in the fundamentals of the subject, as in mastering arithmetic*' (p. 18). Thus diagnostic tests must be based upon a thorough understanding of what is involved in learning, or what has to be learned in order to become proficient. One reason why diagnostic language tests are rare may be because we do not (yet) have an adequate understanding of such processes or components.

Schonell and Schonell argue that results from diagnostic tests differ from those of attainment tests in that the latter are largely quantitative, whereas the former are qualitative or analytic in nature. For diagnostic tests it is said that '*in recording results from diagnostic tests the most effective procedure is to enter all test results and errors in a small notebook kept for each dull (*sic*) and/or backward or maladjusted pupil*' (p. 27). They claim that it is errors that provide most information, and they give detailed examples.

Several diagnostic reading tests are described, but they consist in the main of reading words aloud ('test of analysis and synthesis of phonic units'; 'test of directional attack'; 'visual word discrimination test'). Spelling tests are given, but are not presented as diagnostic, and diagnostic tests of arithmetic focus on the 'basic processes' – addition, subtraction, multiplication, division – and on vulgar and decimal fractions and percentages. 'Diagnostic Tests in Money' aim to help teachers discover the 'kinds of errors made by pupils in applying their basic number knowledge to operations involving money'.

Thus what seems crucial here is to classify errors, and to assess the ability to apply *knowledge* to particular *operations*. A common feature of the tests presented is that they contain a large number of items testing

the same operation, with different exponents (words, numbers), and they are subdivided into the different categories of operation to be tested. The possibility that students could actually learn from a test organized in such a transparent manner cannot be ruled out.

The diagnostic tests of written English that are presented in Schonell and Schonell measure English usage, capital letters and punctuation, vocabulary and sentence structure. The authors claim that these are those aspects which admit of objective measurement. The test of usage is clearly a test of grammar (e.g. *The men s . . . the accident.* Answer *saw. Sally has torn/tore her dress.*). The test of punctuation is an error identification and correction test: (e.g. *peter jones is absent.* Answer *Peter Jones is absent.*). The test of vocabulary is one of semantic relations (e.g. A RUDDER is part of a sledge/boat/shop/bus. COURAGEOUS means brave/strong/sturdy/stupid.). The tests of sentence structure are sentence-combining tasks (e.g. *Snow was falling fast. Jim drove the cows into their shed. Jim left the farm.* Answer *Before leaving the farm Jim drove the cows into their sheds, for snow was falling fast.* or *As snow was falling fast, Jim drove the cows into their sheds before he left the farm.*).

Interestingly, Schonell and Schonell also present tests of composition, on topics like 'What I do on Saturdays and Sundays', as diagnostic tests, and norms are presented for scores on the tests, in terms of 'equivalent English ages'. Sample compositions written by children at various ages are presented, albeit without discussion or explanation. Yet Schonell and Schonell also claim that, unlike most tests, diagnostic tests are applied without time limits – a suggestion not encountered in the language testing literature.

A more recent pamphlet (Dolan and Bell, n.d.), bearing the same title as Schonell and Schonell, is essentially merely a guide to educational testing in general, dealing with matters like reasons for giving tests, standardization, validity and reliability and factors influencing them, as well as test construction. Diagnostic testing is dealt with in ten pages. However, they acknowledge the complexity of diagnosis:

> *Learning is a complicated process which may take many forms, from simple associative learning such as knowing that fire is hot, to complex rule learning such as advanced statistics and problem solving. Many factors such as motivation, previous experience, maturity and alertness bear on performance. At the same time, sophisticated mental processes such as perception, information processing, memory, cognition and language are involved. None of these influences act in isolation; none of them alone will therefore 'cause' learning failure. Thus diagnosis of learning failure is far from simple and involves at least an assessment of the individual's approach to learning and an examination of the efficiency of the mental processes which are thought to be involved.*
>
> *However, not all diagnostic testing is as complicated and involved as the above. It may be that the testing is of a well known scale of learning such as*

> *the progressive grasp of number subtraction, or a diagnosis of a physical attribute such as 'colour blindness'. In these cases there are often fairly precise measures available.* (p. 24)

They then go on to give references to a number of diagnostic procedures. Screening tests are used to identify individuals who are likely to experience learning difficulties, and such tests may even consist of checklists. The authors emphasize the importance of content validity – whether the checklist samples the factors that do actually predict failure – and of the need for test instructions to make clear to the tester exactly how to identify the behaviour being examined. Predictive validity is hard to establish, since remedial action is usually taken to remove the problem. Importantly, they point out that many so-called weaknesses may be normal inconsistencies in development which 'are capable of spontaneous remission' – i.e. which will disappear as the ability develops naturally.

Diagnostic attainment tests focus on specific issues, for example, capital letters. They caution that the reliability of diagnostic tests may be low if there are few items on the scale of ability being measured – the reliability of the test as a whole may be satisfactory, but if scores are reported on subscales, the reliability may be inadequate. They also point out that since many diagnostic tests are intended to identify weaknesses rather than strengths, high scores will not necessarily indicate high proficiency.

Steps in diagnosis are briefly described in terms of identifying and defining the problem, identifying symptoms, measuring and checking, prescribing and evaluating.

As Gardner (n.d.) points out, '*locating those factors which contribute to success or failure in reading* (or any other trait – my comment) *at a defined level requires a precise knowledge of the reading process and care in design*'. He does, however, caution against interpreting correlations of factors in reading with reading performance as indicating cause. He points out that identifying symptoms of a problem does not explain the problem, and failure in learning is likely to be explained by reference to several, and possibly interacting, 'causes'.

Diagnosis of reading difficulties exemplified: Boder and Clay

Before examining in detail the nature of the diagnosis of foreign language proficiency in later chapters, it will be helpful to look at the sorts of approaches that have been developed for the diagnosis of first language reading problems. I could have chosen any number of reading and reading readiness tests, but I have selected two that are different from each other and which both claim to be diagnostic. One is the Boder test, and the other the Diagnostic Survey, both from roughly the same time period (1982 and 1979).

The Boder test

The Boder test of reading-spelling patterns (Boder and Jarrico, 1982) is intended to enable the early diagnosis and remediation of dyslexia for clinical and research use. Patterns of reading and spelling are said to be frequently indicative of cognitive disabilities that relate to brain functions. The test is based on a typology of developmental dyslexia, and is intended to enable what is called direct diagnosis – the analysis and classification of so-called dyslexic errors in reading and spelling performances and relating them to deficit functions. However, it is acknowledged that different people making the same kind of error may well do so for quite different reasons. Quite how this is dealt with is unclear.

In fact, the test is reportedly intended as a screening device for the early identification of dyslexia and other more specific diagnostic tests (unnamed) are available for more detailed diagnoses.

The diagnostic purpose of the tests is

1 to differentiate specific reading disability or developmental dyslexia from non-specific reading disability, through reading and spelling performances;
2 to classify dyslexic readers into one of three subtypes;
3 to provide guidelines for the remediation of all four subtypes (the non-specific and the three specific subtypes).

The test results in five profiles: normal; non-specific reading disability; dysphonetic; dyseidetic; mixed.

The Manual contains a detailed account of each subtype and its characteristic features, and a separate chapter on remediation contains a considerable amount of detail on treatment and procedures. The Manual also has a chapter on the theoretical and empirical background to diagnosis, and an account of the development of the test. In addition, there is a chapter on the reliability and validity of the test, presenting research evidence (of inter-rater reliability, test–retest, internal consistency, criterion-related validity and construct-related validity). There is also a set of Guidelines for identifying Good Phonetic Equivalents (GPEs).

The test has two components, a Reading Test and a Spelling Test. Both are based on standard graded word lists as used in basal readers in Los Angeles schools. A basic construct is the notion of the reading–spelling discrepancy: good readers are said to be able to spell accurately between 70 per cent and 100 per cent of their 'known words'. Poor readers can frequently spell less than 50 per cent of their known words. A Known Word is a word which, presented singly, without context, on a card, can be read aloud instantly – within 1 second. This is also known as 'Flash' reading. An Unknown Word is either one which the student does not attempt, or does not read accurately even after 10 seconds (known as untimed presentation).

The Reading Test consists of 15 lists of words: List A and List B for Pre-Primer readers, and Lists 1–13 for Grade 1 to Adult (Lists 10–12 are High School). Each list contains 20 words, half of which are so-called phonetic words and half are non-phonetic words (a phonetic word is one that follows regular rules of orthography/pronunciation relations).

The test is administered one to one, with the administrator sitting side-by-side with the student. It takes approximately 30 minutes. The instructions are: *Read the words aloud as quickly as you can, but do not hurry.* The administrator either starts at the first level, or at the level at which s/he judges a student will have no problems, and then works up the levels. A student's reading level is the highest grade level at which the student reads less than 50 per cent of the words 'Flash'.

All misreadings are recorded, and are classified into three categories: wild guesses (*diesel* for *doubt*), gestalt substitutions (*sound* for *should*) and semantic substitutions (*ears* for *eyes*). Non-words and nonsense words are also recorded.

Non-word misreadings are classified into two categories: phonetically inaccurate misreading of a phonetic word (*herit* for *with*) and misplaced stress in a phonetic word (usually not corrected, e.g. *tru-ánt* for *truant*).

The reading test aims to identify the student's sight vocabulary, i.e. words read instantly (within 1 second) without being sounded out – this is said to involve whole word processing, or a gestalt approach. The untimed presentation is said to reveal phonic word analysis skills – the ability to read words which are not in one's sight vocabulary.

In the spelling component, the administrator prepares an individualized spelling list, of ten of the student's Known Words and ten Unknown Words (as emerged during the Reading Test), typically doing this whilst the student is doing another task, such as the Alphabet Task (two columns of letters to be matched, one in upper case, the other in lower case) or the Draw the Face of a Clock task. Known Words are selected from those which were read at the student's reading level, actual grade level, or lower, and Unknown Words are selected at the student's reading level or higher.

Once the words have been selected, they are dictated by the administrator, first in isolation, then in a sentence, then again in isolation (*big*; *the house is big*; *big*). The (prescribed) sentences are contained in a separate examiner's folder.

Unknown Words are not given in a sentence. The administrator is required to say: *I do not expect you to spell these words correctly. I just want you to try to write them the way they sound.* The student is asked to repeat the word before writing it down. The test is reported to correlate well with performance on the WRAT – Wide Range Achievement Test – which is a standardized test of reading and spelling which determines reading levels. It is advised that the Boder test, which also determines reading levels, and identifies what students have and have not learned, should not be used alone but within a multidisciplinary programme.

The Clay Diagnostic Survey

A somewhat different approach is presented by Clay (1979) in her Diagnostic Survey. The survey is intended to aid with the early detection of reading difficulties, and is intimately linked with a so-called Reading Recovery Programme. The diagnostic procedures were developed alongside a careful empirical study which aimed to describe 'the range and variability of reading behaviours in children with marked difficulty in beginning reading' (p. 67). It also involved a detailed observation of experienced teachers in order to describe the variability of teaching responses to those children in individual reading tuition sessions. It is worth quoting at length from the account of how the diagnostic and recovery procedures were developed:

> *The difficulties which were demonstrated by the 6-year-old children who were referred to the programme were diverse. No two children had the same problem. Procedures for dealing with these problems were evolved by observing teachers at work, challenging, discussing and consulting in an effort to link teacher and pupil behaviours with theory about the reading process.*
>
> *A large number of techniques were piloted, observed, discussed, argued over, written up, modified, and related to theories of learning to read. Some were expanded, others were discarded . . . Tutor-teachers were challenged by their colleagues, acting as observers, to explain why they chose a technique, a particular book or a specific progression. Tutor-teachers were asked*
>
> - *what contributed to the decision*
> - *how they could justify it*
> - *what other difficulties or achievements the new procedure related to*
> - *why they did not move the child more quickly (out of the recovery programme)*
> - *why the child reacted in this or that way.*
>
> *During such discussions the implicit assumptions of tutor-teachers' decisions were explained verbally rather than remaining intuitive hunches. The process of articulating the basis for a teaching decision was always difficult and sometimes uncomfortable or embarrassing. The procedures arose from the responses of experienced teachers to children's reading behaviours. The process of evolution and refinement continued over three years and the written accounts of these were edited and revised many times. Many techniques were tried and only the most effective remained.* (p. 67–8)

Thus diagnosis, which was always conducted on a one-to-one basis, proceeded alongside remedial teaching, and was almost entirely based on the behaviours exhibited by the children. Although Clay says that it is also possible to make use of standard tests of reading difficulties, like the Neale Analysis of Reading Ability (Neale, 1958), or word tests like the Schonell R1, most of the procedures in the Diagnostic Survey were developed by the Reading Recovery Project, and were only retained if they had the potential to lead to useful teaching procedures. Clay stresses that it is important to use a wide range of observational tasks, since

'*no one technique is satisfactory on its own*' (p. 10). The techniques illustrated in Clay (1979) include a Record of Reading Behaviour on Books, Letter Identification, Concepts about Print and a Word Test using, not standardized tests, but lists of words from the high frequency words in the reading materials being used in class. In addition, examples of children's writing behaviours are collected and analysed, and a locally produced test of writing vocabulary and a dictation test are also included in the test battery. Above all, the Survey aims to record the child's strategies for dealing with texts, words and letters, and is more concerned with the operations the child engages in than with test scores *per se*.

The Record of Reading Behaviour on Books aims to gather a 'running record' of text reading (similar to Goodman and Burke's miscue analysis, 1972), on materials at three levels of difficulty – the child's current book or a selection from that book read at 90 to 100 per cent accuracy, a harder text and an easier text. The running record (usually a sample of 100–200 words from each text read), which takes about 10 minutes to complete, contains a record of every word read correctly and every error made (detailed instructions are given for the recording of errors). Analysis of the reading record includes noting directional movement (starting top left?, left to right orientation?, locating the first letter in words and at the beginning of lines?, etc.), identifying what clues are used, whether the child monitors and self-corrects, and so on.

The Letter Identification measure aims to identify the child's preferred mode of identifying letters (an alphabet name, a sound that is acceptable for the letter, or a response which says 'it begins like ...', and then gives a word for which that letter is the initial letter), as well as the letters the child confuses, and unknown letters. Again, detailed instructions are given for the task.

The Concepts about Print test seeks to explore the child's mastery of significant concepts of printed language, and takes about 5 to 10 minutes. Some of the concepts checked are: which is the front of a book?, that print and not a picture tells the story, what a letter is, what a word is, what is the first letter in a word, the difference between big and little letters, the function of the space, uses of punctuation, and so on. There are 24 items, and detailed instructions are given for the administration of each item. For example:

Item 1: Test: For orientation of book. Pass the booklet to the child holding the book vertically by outside edge.
Say: *Show me the front of this book.*
Score: 1 point for correct response.

Item 15: Test: Meaning of a question mark.
Say: *What's this for?* Point to or trace the question mark with a finger or pencil.
Score: 1 point for explanation of function or name.

Item 19: Test: Capital and lower-case correspondence.
Say: *Find a little letter like this.* Point to capital T and demonstrate by pointing to Tt if the child does not succeed.
Say: *Find a little letter like this.* Point to capital M, H in turn.
Score: 1 point if both Mm and Hh are located.

The Word Test, of only 15 words compiled from the 45 most frequent words in the 12 books of a specific Ready to Read series, is specifically targeted at 6-year-old children who have only had one year of reading instruction. It is emphasized that the Word Test does not give a reading age, nor does it discriminate between better readers after one year of instruction – rather it groups them together.

Teachers are also advised to examine samples of the child's writing behaviour, to see whether the child has good letter formation, how many letter forms s/he has, and whether s/he has a stock of words which can be constructed from memory with the letters correctly sequenced. It is argued that '*a poor writing vocabulary may indicate that, despite all his efforts to read, a child is in fact taking very little notice of visual differences in print*' (p. 22).

In the test of writing vocabulary, the child is encouraged to write down all the words s/he knows how to write, starting with his/her own name. In the dictation test, using simple sentences like 'I have a big dog at home', the child is given credit for every sound that s/he writes correctly, even though the word may not be correct. It is claimed that '*the scores give some indication of the child's ability to analyse the word s/he hears or says and to find some way of recording the sound s/he hears as letters*' (p. 24).

A substantial part of the book is devoted to describing the organization and content of a Reading Recovery programme, giving advice on the selection of suitable texts, how to record behaviour, how to note the child's confusions and the presence or absence of self-corrections, how to introduce new materials and how to increase text difficulty. Particular attention is paid to describing a series of teaching procedures that were shown to have worked well with 6-year-old children who had been identified through the Diagnostic Survey as being unable to make satisfactory progress in regular classrooms. These include:

- Learning about Direction
- Locating Responses
- Spatial Layout
- Writing Stories
- Hearing the Sounds in Words
- Cut-up Stories
- Reading Books
- Learning to Look at Print
- Teaching for Operations or Strategies

- Linking Sound Sequence with Letter Sequence
- Teaching for Word Analysis
- Teaching for a Sequencing Problem
- Strong Skills Which Block Learning
- When it is Hard to Remember (names for letters, or a label for a word) and
- Teaching for Phrasing and Fluency.

As will be apparent, the content of most of these procedures relates directly to information that has been acquired through the diagnostic procedures and is intended to address the strengths and weaknesses thereby identified.

Diagnosis summarized

What these approaches to the diagnosis of reading problems have in common is that they are detailed, they are administered on a one-to-one basis, they are firmly based upon experience with readers having difficulties in reading, they have a close relation to a theory of how children begin to learn to read, and they are, at least potentially, linked to detailed procedures for remediation or recovery. In no sense can such tests or assessment procedures be confused with placement procedures. They supply much more detailed information about a child's abilities and behaviour than is normally taken into account in the sort of placement tests that are used, for example, in institutions that teach English as a Foreign Language. Moreover, the very detailed results lead to a profile of abilities, strengths and weaknesses that are specifically addressed in subsequent intensive instruction. Such diagnostic procedures are often carefully targeted at a narrow age range of client, and are certainly not intended to cover a wide gamut of abilities (unlike the typical placement test).

Although based upon performance on texts, they can hardly be said to be authentic or life-like; instead, they are specifically intended to tap component skills and abilities which are known to contribute to fluent reading. They do not assess fluent reading itself, since if somebody can already read fluently, they are not in need of diagnosis. In other words, the tests and procedures are problem-based and focused on identifying behaviour, knowledge or strategies whose mastery is essential to the development of an integrated, meaning-based approach to the understanding of a wide range of written texts.

Diagnosis in first language versus diagnosis of foreign language

It seems to be the case that the diagnosis of language, reading or learning difficulties in a person's first language is quite common, yet the diagnosis of foreign language proficiency seems to be infrequent, or at least unreported. Why might this be?

One obvious answer is that an ability to read in one's first language is central to being able to participate in education, in employment and in society more generally, just as disorders in one's first language can have a catastrophic effect on one's ability to communicate as a human being. Thus diagnosis is most developed and best understood in circumstances where the skills that are lacking or in need of remediation are important to life, education and society, and where diagnosis can contribute to the quality of life of individuals. Learning to use a second or foreign language may be important, especially in multilingual societies or where the learner might be significantly disadvantaged by a lack of proficiency in the language of the community. But it is rarely the case that one's life chances would be significantly impeded by a lack of foreign language proficiency. Learning to read is something that is expected of every child in modern societies, and an inability to read is seen as a major handicap. That is not the case for second or foreign language learning.

Note, however, that I am not claiming that it is easier to learn to read than to learn a foreign language. Indeed, one could argue that although it is achieved by most members of modern societies, the process of reading and the process of learning to read is very complex, and not well understood, despite centuries of research. The very importance of being able to read has resulted in a great deal of attention being devoted to helping those who have reading difficulties, and it is partly because of this attention that ideas about diagnosis have developed. As the Reading Recovery Programme shows, paying detailed attention to the behaviours that are exhibited (and not exhibited) by children who are experiencing difficulties in reading is both important and productive. There is a great lack of similar detailed attention to the difficulties experienced by learners of a foreign language, and there can be little doubt that an increase in the volume of descriptions of such difficulties, alongside advances in second and foreign language acquisition and learning theory, would pay dividends in terms of an enhanced ability to diagnose strengths and weaknesses in foreign language learning. Recall Dolan and Bell's statement, cited above: '*Diagnosis of learning failure is far from simple and involves at least an assessment of the individual's approach to learning and an examination of the efficiency of the mental processes which are thought to be involved*'.

However, there are features of this tradition of diagnosis in one's first language that may be distinguished from current approaches in foreign language education.

First, the focus in these procedures is very much on identifying problems and especially errors. Informants' *mis*readings are recorded, their *failure* to identify letters is noted, their *mis*understanding of print and their *lack* of readiness to read are in focus. Orthodoxy in foreign language education has moved away from problems and errors to strategies for overcoming problems. Traditional diagnosis focuses on behaviours, whereas foreign language education tends to emphasize the cognitive

aspects of foreign language learning. Traditional diagnosis is often based upon a typology of problems, behaviours and typical difficulties of those who are in some sense failing. The emphasis is on abnormal behaviour and manifestations of ability, rather than on the normal.

This is very far from current orthodoxy in language education, where failure is not a commonly used concept, and where the emphasis is on the normal or the supernormal ('The Good Language Learner') rather than on what we can learn from those who fail to learn. It could well be that diagnosis of foreign language proficiency would benefit from paying more attention to causes of failure in future.

Nevertheless, there are clearly also points of similarity. The recognition that the abilities being addressed and their underlying features and causes are complex is certainly a common feature of discussions in foreign language education and the important role of background, motivation, memory, cognition and perception is equally acknowledged in foreign language education. As is the complex interaction among such factors in contributing to one's development in a foreign language and one's ultimate success.

Similarly, the importance of understanding the mental processes involved in learning is stressed in foreign language education, as is the need to understand what is involved in normal language use, be that using a foreign language or reading in one's first language.

Moreover, modern foreign language education also stresses the need for qualitative information about learners and learning, and emphasizes the need for data that is not just test-based, but which is based upon observation of behaviour. The importance of a detailed record of a learner's responses would also surely strike a chord among researchers of foreign language learning in the twenty-first century. The statement that no one technique of diagnosis on its own can adequately identify a child's reading problems would also be echoed by methodologists of research into applied linguistics.

Thus there are many points of similarity in these two superficially different fields, which should encourage those who are interested in developing better diagnoses of foreign language proficiency. And therefore it would be rewarding for language testers to re-examine their current prejudices and practices in assessment, and see whether they might not have something to learn from older and more traditional practices in diagnosis.

Attention in the twentieth century has largely focused on assessing one's ability to communicate in a foreign language, on conducting analyses of learners' needs for the language in the real world and on analysing Target Language Use situations, and seeking to simulate as closely as possible those settings and uses. Authenticity has been a byword, in various disguises. The Common European Framework itself is the epitome of such attempts, viewing language in social action. Language testing has tended to downplay the importance of language itself and above

all the role of language knowledge in language use, much as the Common European Framework does not address the specific issues of specific languages in dealing with communicative competence. It could well be that diagnostic testing of foreign language proficiency would benefit from considering more closely issues to do with language and failure to learn specific features of language when developing its test constructs.

Nevertheless, it is clear that both traditional and current diagnostic testing need to pay attention to the construct they are attempting to assess. Without a theory of development, a theory, perhaps also, of failure, and an adequate understanding of what underlies normal development as well as what causes abnormal development or lack of development, adequate diagnosis is unlikely. For practical reasons it is unlikely that foreign language diagnosis will involve individual, detailed, one-to-one assessment of the sort we have seen exemplified above, although there is in principle no reason why a computer could not deliver an individualized and detailed set of elicitations to learners, and store the results for detailed analysis and profiling by humans. (It is noteworthy that what was said in the previous chapter about the importance of delivering a large number of parallel 'items' in order to gain a representative, reliable and reasonably complete picture of a learner's ability is paralleled in the discussions of diagnosis in the examples presented in this chapter. Such a large battery of items is arguably best administered by computer.)

In the present context also, the focus is on how diagnostic *tests* can be developed, and not other diagnostic assessment procedures, although I freely acknowledge that diagnosis by tests is inevitably limited in a number of ways. My point is rather that decades of emphasis on learner-centred teaching, on individualized instruction, on self-assessment and the importance of learner awareness have failed to come up with diagnostic procedures that have general applicability and value based on one-to-one, individualized procedures for diagnosis. Indeed, I am uncertain whether the issue and the need for diagnosis have even been addressed. The point I wish to make throughout this book is that they should be addressed.

However, until such time as much more research is undertaken to enhance our understanding of foreign language learning, we will probably be faced with something of a bootstrapping operation. Only through the trial and error of developing diagnostic instruments, based on both theory and experience of foreign language learning, are we likely to make progress in understanding how to diagnose, and what to diagnose.

Chapter 3: Introduction to DIALANG

Issues in diagnostic testing

So far we have seen that, in contrast to the field of first language assessment, in second and foreign language testing and assessment diagnostic testing has received much less attention and discussion, and as a result there is considerable confusion and indeed ignorance about what diagnostic testing might entail. It is clear from the vague and contradictory definitions in the foreign language testing literature, and from the lack of published research into diagnostic tests, as well as from Web-searches, that in the twenty-first century we are still facing the lack of diagnostic tests that Hughes noted in 1989.

Chapter 1 revealed that there are many issues concerning the design and use of diagnostic foreign language tests that are touched upon but remain unexamined in the literature. The list below is but a sample of the issues which will need to be addressed if we are to make progress in developing useful diagnostic tests:

- Is placement really the same thing as diagnosis?
- Should diagnostic tests be global in nature or engage in more detailed testing?
- Why are diagnostic *tests* rare?
- Are diagnostic *uses* of tests common, as has been asserted, and, if so, in what sense are such uses diagnostic?
- How, indeed, are diagnostic tests used?
- How are other tests used for diagnostic purposes, and are such uses justified and validated?
- What should diagnostic tests test – learner strengths or learner weaknesses?
- What diagnosis is useful and in which settings?
- What skill, subskill or linguistic feature is susceptible to treatment, since if remediation is not possible, there may be no point in conducting a diagnosis in the first place?

- How should the results of diagnostic tests be reported? Merely as an overall, global score? A profile of scores across skills, or subskill by subskill? A detailed report, linguistic feature by linguistic feature (present perfect, third conditional, and so on)?
- How, finally, should the interpretation of the results of diagnostic tests be validated?

In the light of such unanswered, indeed so far unaddressed, questions, I am therefore tempted to suggest that it is not that diagnostic tests would not be useful, rather it is that we do not yet know either how to write them – what they should contain – or how to interpret the results.

The reader may well ask whether such a conclusion is justified. Surely we ought to be able to identify linguistic features – structures, words, phrases, speech acts, functions and notions, skills and microskills, strategies and aptitudes – that language learners need to acquire? After all, these are contained in many textbooks and syllabuses, they are written about in applied linguistic journals and in handbooks for teachers, and they are presented in papers at conferences. So why do we not have tests to diagnose whether somebody does or does not yet master these various features?

Partly, no doubt, this is because nobody pays them to do so. Most tests are developed for specific purposes, and these are often high-stakes, important purposes like school-leaving exams, university entrance tests, international proficiency tests, and the like. Institutions often have a specific use for placement tests, or for tests based upon a specific textbook, but apparently not for diagnostic tests. Increasingly, publishers produce tests to go along with their materials – although many of these are rather weak and not representative of the variety of content and exercise types contained in the same textbooks. Publishers even sell placement tests to place students onto courses suitable for their level.

Yet evidence is rarely presented on how such proficiency or placement tests perform which would enable us as language testing researchers or applied linguists to study what students have problems with, what they can indeed master, and what we can learn about the nature and development of language proficiency as a result of such tests. This may be for commercial reasons: considerable investment goes into producing such tests, and it might give commercial advantage to others if the results were widely known. The lack of available data may also be because the data has never been analysed in a suitable manner, or it may simply be that the data is not as helpful as one might wish it to be.

Another reason why diagnostic tests have not been developed may be because there has not until recently been a suitable framework of language proficiency which attempts to describe how learners' language proficiency develops. Such a framework would describe what it is that distinguishes an advanced student from an upper intermediate student,

what the difference is between a false beginner and a true beginner, or a lower intermediate student. Teachers, textbook writers and test developers doubtless all have their opinions on such matters, based upon their own experiences, but it is an open question as to whether these experiences hold true for learners whose experiences are different from their own, whether they are generalizable, and how they relate to what we know about language proficiency in general.

However, frameworks of language proficiency have been developed in some contexts, including, in Australia, the International Second Language Proficiency Ratings (Wylie and Ingram, 1995/99), in the USA the ACTFL scales (American Council for the Teaching of Foreign Languages, 1983), and the Canadian Benchmarks (Pawlikowska-Smith, 2000), all of which attempt to describe language proficiency. In Europe, the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) has recently been developed (Council of Europe, 2001). This is actually a synthesis of the Council of Europe's work over some 30 years, from the notional-functional syllabus to the Threshold Level for English, through versions for a large number of other (mainly European) languages, then to Waystage, Vantage, and finally the full six levels of the CEFR. Thus we now have a number of theoretical frameworks of language use, language learning and language development which could provide a basis for the construction of instruments that could diagnose what level a learner has reached, what that learner's strengths and weaknesses are, and how that person might best be advised to improve and build on their language proficiency. In short, the frameworks are now there: the question is whether diagnostic tests based on such frameworks will follow.

Interestingly, one of the precursors of the CEFR, the Threshold Level, was and is very popular. It was the basis, for example, of major language programmes like the BBC's *Follow Me* series, and tests were developed in parallel to the TV programmes and textbooks. Yet, to my knowledge, no reports are available that give information about what learners found difficult or easy on such tests, nor is there any evidence for what learners who could be deemed to be at the Threshold Level could actually do on tests designed to measure their ability and their performance. This would seem to be a missed opportunity or, at least, an avenue in language testing history to be followed in case there is indeed evidence available somewhere that might throw some light on language development and thus on possible bases for diagnosis.

In point of fact, the University of Cambridge Local Examinations Syndicate (UCLES) has developed a suite of tests, which they call their Main Suite, at five of the six levels of the CEFR scale. The tests are the Key English Test (KET), the Preliminary English Test (PET), the First Certificate in English (FCE), the Certificate of Advanced English (CAE) and the Certificate of Proficiency in English (CPE). Information on how items on these various tests perform could be extremely useful

in developing a theory of what distinguishes learners at level X from learners at level Y. To the best of my knowledge, however, no such information has been published. It would be very rewarding to examine such tests for their diagnostic potential, or at least for the light they might throw on what changes as learners become more able, as they progress through the different levels and succeed on the different Cambridge tests.

However, one project has been explicitly based upon the CEFR, and has attempted to develop test items for five different aspects of language proficiency at each of the six CEFR levels. This project was funded from public money and has not yet imposed any restriction on the analysis of the data that result from such tests. This project is known as the DIALANG Project, and research is currently ongoing into the results of that Project. It is to be hoped that, as a result of that research, we might begin to learn more about the development of language proficiency which could help us better to diagnose learners' ability. This book aims to explore how the DIALANG Project, its assessment procedures and the data that are available from the application of those procedures might contribute to a better understanding of diagnosis and the diagnostic potential of tests and assessment instruments.

It is precisely because a project like DIALANG exists that we must seek to study what it was thought important to include in a diagnostic test battery, to explore what did actually work, what proved useful and what less useful, and how individuals can be described as developing in their language proficiency. Once we know what develops and in what sequence, at least for one language, then perhaps we will be in a better position to diagnose what has not developed, or to understand how to help an individual to develop further, which is, after all, the purpose of diagnosis.

The DIALANG Project

The DIALANG Project was set up explicitly to develop diagnostic tests, not proficiency tests, in 14 European languages. The test are delivered over the Internet, and thus, although the item types are necessarily somewhat limited, this means that it has been possible to gather information from a fairly large number of candidates, at least for English, from a wide range of national and linguistic as well as educational backgrounds.

In addition, not only has background information been gathered about the candidates – what their mother tongue is, how long they have been learning English, and so on – we also have their own assessment of their language ability, using both the global and the detailed skills scales from the CEFR (modified by DIALANG, translated into all 14 languages, and published in the appendices to the 2001 version of the CEFR). We can therefore explore the relationship between learner performances on the test tasks, and their self-assessments.

DIALANG has tests in five aspects of language and language use: Reading, Listening, (indirect) Writing, Grammar and Vocabulary, for all 14 languages. Results on the six levels of the CEFR are reported on each of these tests separately, and are not aggregated into an overall result. The individual items are designed to test particular aspects of the skill in question. Thus for Reading, for example, there are items that test the ability to distinguish the main idea from supporting detail, to understand the text literally, and to make appropriate inferences. Results on these 'subskills' are reported separately, so that students can see their responses to each item, and also on each subskill generally, as items are grouped according to the skill they (supposedly) test.

Learners also self-assess their abilities in Reading, Listening or Writing. They are given extensive feedback of two sorts, in addition to the detailed test results. First, they are given feedback on the difference between their self-assessment and their test performance and they are invited to explore reasons why such discrepancies may occur. Such information is intended to raise learners' awareness about the nature of language and language learning, and thus to help them diagnose for themselves what their strengths and weaknesses might be in the light of such feedback.

Second, they are given what DIALANG calls 'advisory feedback', which is in essence advice on how learners might progress from their current level to the next level up on the CEFR. The aim of such advisory feedback is to encourage learners to think about what might be an appropriate action if they wish to improve their language proficiency. In short, the DIALANG system aims to encourage self-diagnosis and self-help.

It is claimed that DIALANG is diagnostic in a number of different ways. First, it is diagnostic at the macro level, in terms of performance overall in the skill being tested and how a learner's performance relates to the levels of the CEFR. This can be used, for example, to help learners or their teachers decide whether they are ready to take a public or commercial examination at that level, or whether they might be better advised either to improve their language or to take an exam at a lower level. That is clearly a useful, albeit fairly general, level of diagnosis. Information at this overall level can also be used for placement purposes.

But the tests are also diagnostic at a more micro level, in terms of the subskills that are being tested. Thus, learners may discover that whereas they are good at understanding texts literally, they may be weaker at making suitable inferences. Such information may be useful in helping them, or their teachers, decide what to concentrate on next in language learning or teaching.

And finally the tests are diagnostic in the sense of self-diagnosis: as explained above, learners are, through the provision of verbal feedback (presented in all 14 languages of the system), encouraged to reflect upon

their performance, how it compares with the CEFR and their own beliefs and expectations about their ability, and they are by implication encouraged to take whatever action they deem appropriate.

What is noteworthy about this system is not only that it is free of charge, but also that it is not a high-stakes test. Indeed it is not even a low-stakes test. It is a test with no stakes. DIALANG does not offer certificates, and is entirely learner-centred. DIALANG, therefore, does not compete with any other test. There are many tests that offer certificates and that can be used to obtain a job or to enter university. Since DIALANG is 'no-stakes' there is therefore no need for elaborate security, and because the test system is freely available there is no need for a registration procedure or even collection and storage of results (although tailor-made versions can indeed store and analyse results). The tests are aimed only at the learner him- or herself. If learners wish to tell their teachers, friends, colleagues or bosses what level they have achieved and how they have been diagnosed, that is entirely up to them.

Many who learn about DIALANG suggest that it is easy to cheat on such a test: all one has to do is to write down the items, learn the correct answers and then go back into the system and get a higher score. That is entirely possible. But anybody doing so would only be cheating themselves.

This is possibly one example where a diagnostic language test is similar to a medical diagnosis, despite the difference in stakes involved. Who would want to cheat on a medical diagnostic test? You might not like the result, of course, but what would cheating achieve? The good news about taking DIALANG is that there are no life-threatening consequences.

In this book, I shall refer to the DIALANG Project frequently, and therefore in the remainder of this chapter I shall describe the DIALANG system in some detail. Readers who are interested in seeing for themselves what DIALANG is like can download the system and associated tests from www.dialang.org, and experimental item types associated with the DIALANG Project are presented in Chapter 15, and are also available on the DIALANG website.

A description of DIALANG

DIALANG is an on-line diagnostic language testing system, that contains tests of five language skills or aspects of language knowledge (Reading, Listening, Writing, Vocabulary and Grammar, also known as Structures) in 14 European languages. These languages are Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish. The test specifications are based on the Common European Framework (CEFR), and the results are reported on the six levels of the CEFR (A1 – lowest – to C2 – highest). The interface of the system, or what the Project refers to

as the ALS (Administration Language System) – test instructions, controls, help pages, explanations, self-assessment statements, test results, feedback and advice on how to improve one's language proficiency – is also available in these 14 languages.

Users have complete freedom to choose which language they wish to see instructions in, which language they wish to be tested in, which skill they wish to be tested in, whether to take an initial vocabulary test that will estimate their ability, whether to answer a series of self-assessment statements about their abilities in reading, listening and writing, and which results they wish to examine. They are also free to read feedback on their performance and advice on how to progress, they can choose to take as many tests as they wish, and they can quit a test at any point. The system is intended to be maximally flexible (14 ALS options × 14 test languages × 5 skills at minimally 3 levels of difficulty), and it is free of charge to users.

The system is currently downloadable (without charge) via the Project website at www.dialang.org. Since the download includes Java programs, this can take some time if one is using a 56K modem, but downloads over broadband are rapid. Elements of the program are stored on the user's computer, but users have to be connected to the Internet to take the tests, since test items and screens are sent from the Project servers to be cached on the user's machine at the beginning of a test session, and user responses are sent back to the Project servers during the actual session.

Below I describe how the system works, so that subsequent chapters are more readily understood. However, readers can download the program and explore it at their leisure.

When users enter the DIALANG system after going online to the Internet, and clicking on the DIALANG icon on their desktop, they first see a Welcome screen, and are offered the choice of the interface (the ALS) in any of 14 languages. Once they have selected the ALS, all subsequent information is presented in the language chosen (except, of course, for the test items themselves).

Next, they are presented with a brief description of the DIALANG system, and they can scroll through screens that describe the different phases of the testing system, or they can skip these and go directly to the Test Selection Screen.

Before users can choose which language they wish to be tested on, and which skill, they are presented with a disclaimer which advises against test misuse, since these tests are intended to be diagnostic and not certification tests. Once users have indicated that they understand this test misuse warning by clicking on the OK button, they can proceed to select their language and skill.

Depending upon the language chosen, another disclaimer may warn the user that the test chosen is still being piloted in order to determine the difficulty of the items, and that therefore the test results are currently

based upon expert judgements of item and test difficulty, until such time as calibrated data are available. Once a language has achieved sufficient numbers for dependable statistical data to be available, this warning notice is removed.

Having selected a language and skill, users are asked to confirm their choice, and are then presented with a screen explaining the Vocabulary Size Placement Test (VSPT) and its function. The VSPT is a Yes–No vocabulary test which is used in Version 1 of DIALANG in order to assist in the decision as to which one of three levels of difficulty to administer tests at. The test consists of a list of 75 words, and the user's task is to say whether each word is a real word in the test language or not. Fifty of these words are real words in the language, 25 are pseudo- or non-words. The user is not told how many pseudo-words there are, and all words are presented in a randomized order. The user receives one point for each word correctly identified as real or pseudo – i.e., a response 'Yes' to a real word and a response 'No' to a pseudo-word.

Once users have completed the VSPT, they are given immediate feedback on their performance in terms of six different bands of ability, ranging from *very low* to *indistinguishable from a native speaker*.

Users can, however, choose not to take the VSPT, or to quit whilst in the middle of the test. If they do so, they are given a warning of the consequences, namely that the test they subsequently take may be too difficult or too easy for them, since users' performance on the VSPT contributes to the selection of a skill test of suitable difficulty.

On completing or skipping the VSPT, users are next offered the opportunity of self-assessing their ability in the language and skill they are to be tested on (the self-assessment statements SAs – are in the ALS). These statements are couched in the form of 'I can ... (do X)'.

There are self-assessment statements for reading, writing and listening, but not for vocabulary or grammar, since the CEFR itself does not contain any language-specific self-assessment statements.

Users are requested to read each statement in turn (there are 18 for each skill), and simply to respond 'Yes' if they believe that they can do what is described in the statement and 'No' if they cannot. The statements are arranged in order of increasing difficulty. Once users have responded to all 18 statements, the system calculates the user's CEFR level, basing its estimate on the logit scale. Once users have taken the language test proper, the result of the self-assessment is reported and compared with the user's test result. The system then offers the possibility of explanations for any discrepancy (see below). As with the VSPT, users can skip or quit the self-assessment component, but are warned of the consequences if they do so.

If users have responded to both the VSPT and the self-assessments, then the two results are combined to decide which level of test they will be given. At present, there are three levels of difficulty of test – 'easy', 'medium' or 'difficult'. (However, the program can in principle

accommodate any number of levels of difficulty.) If users have only responded to one of the two placement procedures, then that result alone is used to make the placement decision. If no placement procedure has been selected by the user, then the 'medium' test is administered.

Once the self-assessment procedure has been completed, or skipped, users are presented with a test of the skill and language they have previously chosen. First they see a screen explaining the test, and offering the possibility of receiving immediate feedback, item by item. Then they are presented with the test items, item by item. They can switch immediate feedback on or off throughout the test.

The test items are in one of four different formats: multiple-choice, drop-down menus, text-entry and short-answer questions. For those test items that require some written response, they can select a virtual keyboard in order to input special characters as required by the language they are being tested in.

Users can quit the test at any point but are warned of the consequences of doing so. In the present implementation of Version 1, users cannot return to change their responses to previous items, nor can they scroll through the test and choose which items to respond to first.

Version 1 of DIALANG is in a sense adaptive, because users are given a test in the light of their performance on the VSPT and their assessment of their own ability in the skill/language in question. This is what DIALANG terms 'test-level' adaptivity. However, Version 2 of DIALANG will be adaptive at the item level for those languages for which sufficient responses have been gathered, and for which a suitably large bank of calibrated items is available.

Once users have finished the test, they see a screen which offers them a variety of feedback: four different types of results, and two sets of advice. If users quit the test before completion, then they only see part of the feedback. They see item review (of the items they completed before quitting) and 'About self-assessment', but nothing more.

'Your level' gives users their test result in terms of the six levels of the CEFR – they are not given a score, which would be meaningless – and they are also given a brief description of what learners at that CEFR level can do, derived from the Framework.

'Check your answers' gives the user a graphic presentation of which answers they got right and which they answered incorrectly (red being incorrect and green being correct). The items are grouped according to the subskill that they are intended to test, so that the user gets an immediate overview of which subskill is strongest and which is weakest.

If the users then click on any one of the items, be they red or green, they are taken to the item itself, where they can see (or hear, in the case of Listening) the text if any and read the original question, their response, and the correct response(s). This screen is identical to the screen they could have seen in immediate feedback mode.

'Placement Test' shows users a screen with their score on the VSPT (from 1–1000), and a brief description about what the score means in terms of vocabulary size. This is, of course, not based on the CEFR.

'Self-assessment feedback' takes users to a comparison of their test score (reported in terms of CEFR levels) with their self-assessed CEFR level. It also gives a mild warning about the risks of over- and under-assessment of one's abilities, and offers users the opportunity to find out more about self-assessment.

'About self-assessment' provides a number of reasons why self-assessment and test results may not match, both in general terms on this screen, and in detail under a number of hyperlinks, with headings like

- How often you use the language
- How you use the language
- Situations differ
- Other learners and you
- Other tests and DIALANG

and so on, and each of these pages may have links to more detail.

Finally in the feedback and advice section, users can select 'Advice', and they will then be presented with a series of tables. The table that appears shows some of the differences between the level the user was assigned to, and the next level above and below.

These differences are based upon the CEFR, and address the type of text that can be understood, what can be understood and the conditions and limitations of that understanding. Users can scroll through these, and they can also look at the other levels of the CEFR by pressing the buttons in the toolbar. The aim of this is to encourage learners to reflect on what is involved in language learning, and how their current level compares with other levels, including those below as well as above.

Users can also choose a button which takes them to a set of suggestions that may help them to move from their current level to the next one above. As is usual in DIALANG, users are free to scroll through this advice, seeing what sorts of suggestions are made for progress to other levels also.

Once users wish to move on from Feedback and Advice, they are given a choice of quitting, or of selecting another language or another skill. If they select another skill, they are not presented with another VSPT but, if that skill is Reading, Writing or Listening, they are given the opportunity to assess their own abilities in that skill. If, however, they choose another language, then they are presented with a VSPT in that language. If they wish to change the language in which they are receiving instructions and feedback, then they quit to the very beginning of DIALANG, where they may exit the program, or choose another ALS. Users also have the option to quit and change ALS at any point in the program.

Chapter 4: The history of DIALANG

DIALANG was born out of a White Paper published by the European Commission in 1995: *Teaching and Learning – Towards the Learning Society* (http://europa.eu.int/comm/education/doc/official/keydoc/lb-en.pdf). The vision contained in this paper included far-reaching objectives for the recognition of skills in all areas of education and training, including through the use of new technologies. DIALANG was the response for the field of language skills, and was part of a family of projects covering a wide range of subject areas, including Mathematics, Chemistry and ICT skills.

The original idea was aimed at encouraging lifelong learning and certification in a Europe that increasingly required its citizens to adapt to changing technologies and employment opportunities. Learning for life – the idea of an education that fitted one for a lifelong career or occupation – had outlived its day, and society needed its workforce and professionals to be able to adapt. Lifelong learning replaced learning for life. Thus individual 'skills passports' would present an opportunity to provide evidence that one had indeed updated one's skills and knowledge. One example of this ideal that has been implemented is the European Computer Driving Licence (see The European Computer Driving Licence Foundation, at http://www.ecdl.com), a sister initiative of DIALANG's, now operating out of Dublin, with a network of national offices. Another offshoot of this general desire to be able to attest to ongoing development in lifelong learning, albeit from a rather different source, is the European Language Portfolio, developed under the auspices of the Council of Europe (http://culture2.coe.int/portfolio/).

The original idea for Europe-wide certification of language proficiency was for a supranational system that would encompass all official languages of the European Union. The system would make use of modern technology – CDs were explicitly mentioned – and would enable certification not only in widely used languages like English, French or German, but also, and of crucial importance in Europe, in

the less widely taught languages of the Union. The framework of DIALANG finally incorporated a system that would include all 11 official languages of the EU, plus ultimately Irish, Icelandic and Norwegian, thanks to the strategic intervention of the national authorities for these languages. Initially it had been hoped to include Letzeburgesch, but no test development team could be located.

Informal discussions led to a modification of the proposal for the development of certification into a proposal for a diagnostic system that would complement existing systems of language certification rather than compete with them. The initiative was taken up by the European Language Council and a project coordination team emerged from among its members. The University of Jyväskylä, which had already designed and was running a multilingual testing system for national purposes, and thus had the experience and personnel needed in this kind of activity, took on the coordinator's role and the Project took off in 1996.

The first meeting to discuss and plan the system took place in February 1997 at the University of Jyväskylä, with representatives of the Centre for Applied Language Studies of the host university, the Free University of Berlin and Lancaster University present. Subsequent meetings in Lancaster and Louvain-la-Neuve in the spring of 1997 and Aix-en-Provence in the autumn of the same year developed the original outline plan into a workplan, with a management structure, a set of test developers per language (known as ADTs – Assessment Development Teams) and a search for partner institutions to support the ADTs. The original proposal and the firm wish of the Commission was that tests be developed simultaneously in all 15 languages, that modern technology be incorporated as far as possible and that a Europe-wide network be developed for the delivery of the tests.

Although initially envisaged as a one-year initiative, early planning made it clear that this was a much more substantial piece of work than originally envisaged and would require far more time than a mere 12 months. The growing partnership realized early on that the development of a Europe-wide system of delivery would be a logistical nightmare, since it would be expected to be independent of any existing, and potentially rival, organization, it would need to be available throughout Europe, and it needed to be set up as soon as possible. Under such conditions, paper-and-pencil tests presented horrendous problems of construction and delivery, and could hardly be said to incorporate modern technology. Even CDs, which were briefly contemplated, presented problems of delivery, and once a CD had been created and distributed, updating and modification of content would be nigh impossible. In addition, test items would need to be piloted – tried out on target populations – in order to establish their quality, and the piloted items would need to be calibrated in terms of their difficulty. Moreover, the very purpose of DIALANG – diagnosis – would have been

extremely laborious to carry out with paper-and-pencil tests, whereas computers enable immediate diagnostic feedback to be given to the users. Ease and speed of providing feedback were therefore key factors in taking the decision to develop a computer-based system.

As early as March 1997 in Lancaster, the Project team discussed such problems, and a tentative decision was made to explore the possibility of delivering pilot tests as well as a live system over the Internet. Lancaster had been developing ideas for computer-based testing for almost a decade (see Alderson, 1986a, 1986b, 1988 and Alderson and Windeatt, 1991), and had already developed prototype systems which were intended to demonstrate the potential of computer-based delivery in the face of growing criticism that computers implied the resurgence of item types such as the cloze and multiple-choice techniques which were rapidly going out of fashion in communicative language testing circles (see Alderson 1986b).

The first such system was developed by Lancaster for the BBC computer at the insistence of the funding body, the British Council, which at the time used BBC-B computers in its English language teaching centres. Predictably, by the time the BBC prototype system had been developed, the British Council had replaced its BBC computers with IBM PCs, and so a further project, funded by Lancaster University and the Innovation in Higher Education Project, developed LUCAS – the Lancaster University Computer-based Assessment System – which was in essence a rewriting of the BBC programs in Visual Basic for the IBM PC.

The DIALANG team asked Lancaster to demonstrate the LUCAS system at its meeting in Louvain in June 1997, and a formal decision was made to contract the LUCAS team to develop the software for DIALANG as the basis of an Internet-delivered assessment system. However, it soon became clear, especially in the light of Lancaster's experience of the rapid development of computer platforms and software, that a future-proof technology should, as far as possible, be implemented. In mid-1997 the only reasonable possibility was Java, a very new programming language. Finding programmers with significant levels of skill in such an immature and rapidly developing language proved difficult and even once the language started to settle it continued to be difficult to get programmers because of the demand. A series of compromises had to be made until a small team could be recruited and designs drawn up for the development of the software.

The rapidly changing Java language was not being matched by changes to the versions of Java used in Web-browsers. As DIALANG required the latest features of this developing language, a decision was taken to move development out of Web-browsers but to keep using Java so that it could be re-integrated into browsers at a later date, when things would (hopefully) have stabilized. The Java program being developed still used the communication methods of the Web to maintain itself as an Internet-based program but did not sit within a Web-browser.

In parallel with these decisions about test delivery, software development and the assembling of ADTs, the Project team had to decide the theoretical basis for the language tests. Given the time line to which the Project was working (delivery was still required within a matter of months), it was clear that there simply was not time, even if it were thought academically desirable and theoretically possible, to develop a common framework for the development of the tests, especially one that could apply to all 14 languages. The Bachman framework, both in its early (Bachman, 1990) and subsequent (Bachman and Palmer, 1996) guises, was a potential candidate for DIALANG, but, as numerous critics pointed out, it was still far from clear how it could be operationalized. It was also very much a product of the USA and of language testing researchers, whereas the clear political as well as practical and even academic imperative was for a model or a framework that could be said to be truly European, and that would engage the adherence of the different assessment cultures in Europe.

The obvious candidate for such a framework was the emerging Common European Framework of Reference for Languages: Teaching, Learning and Assessment (CEFR), being developed under the auspices of the Council of Europe. In 1997, the only version of this framework available was the first draft of 1996, which was in current circulation. Nevertheless, this Framework had such a respectable European pedigree (relating specifically to the widely known and accepted Threshold Levels – Niveau Seuil, Kontaktschwelle, and those translated into numerous European languages from 1975 onwards – as well as to subsequent levels like Waystage and Vantage, which used the Threshold Level as their sources) that the DIALANG Project agreed almost unanimously and immediately to adopt the CEFR as the basis for its own test specifications. The DIALANG Assessment Framework (DAF) and the DIALANG Assessment Specifications (DAS) were thus rapidly developed with reference to the CEFR by the coordinating team in Jyväskylä, and drafts were circulated around the partnership for comment and modification.

As will be remembered, Assessment Development Teams were already being identified, set up and put to work, in the light of the urgent need to develop tests in 15 languages in parallel. In the event, it proved impossible to identify a willing partner for Letzeburgesch, but ADTs were created for the remaining 14 languages and work commenced in haste. The DAS went through six iterations before it was decided that enough had to be enough and, imperfect though it may have been, the sixth version of the DAS was to become the official working document.

At the same time, the LUCAS programming team was rapidly developing templates for test items in a yet-to-emerge authoring system that could be used for the input and review/revision of items across Europe. As with the test delivery system, an authoring system based on paper

and pencil and which relied upon traditional methods of submission and revision, with item writers as far apart as Iceland and Greece, Portugal and Finland, the Isle of Rhum and Sicily, was likely to create another set of logistical problems and so the logical decision was taken fairly early for test development also to proceed via the Internet.

The Project thus faced a chicken-and-egg problem: whether to identify required item types for test construction with regard to the dictates of current language testing theory and practice, which would be within the capacity of ADTs at various stages of expertise and experience to develop, or whether to select item types that would be within the capacity of the still-emerging programming team to develop, within a still-developing modular software system using a still-developing programming language. The eventual compromise was a fairly traditional system of two major item types, with two variants each – multiple-choice and drop-down menus on the one hand, and short-answer questions and text-entry items on the other. Inevitably, some ADTs had already written items which did not conform to these types, or to the developing templates, and these had to be held in store for later implementation, but the vast majority of items produced conformed to these test methods.

The Project was, however, painfully aware that such item types were fairly traditional and did not correspond to the innovative culture of the Project. Nor did they show the potential of a computer-based system for interesting innovation. It was thus decided, still within the first year of the Project, that the programming team, in conjunction with the coordinators of test development, would at some point develop a series of experimental items which would demonstrate to the Project's sponsors and other interested parties what sort of interesting innovations in test method could be developed through computer delivery – in due course (see Chapter 15).

Thus, towards the autumn of 1997, key decisions had been taken about test design and computer delivery systems, teams had been assembled, and work was going apace on writing items for tests of Speaking, Writing, Listening, Reading, Grammar and Vocabulary in 14 European languages. (Tests of Speaking, using indirect methods, were later dropped. The Writing items developed are indirect tests of aspects of writing ability.) Vocabulary size placement tests, to be used to identify a suitable level of difficulty at which the final tests would be administered to clients, were being developed in six of the DIALANG languages. In addition, the self-assessment statements associated with the CEFR were scrutinized, suitable statements were selected, modified to suit DIALANG's purposes and audience, and translated from the English originals into the other 13 languages of the Project. Test instructions – rubrics – were also developed for the pilot tests and translated into the other 13 languages, and feedback was designed.

However, the very tight timetable and the decentralized nature of the item writing involved the coordinating centre in a great deal of training of item writers and continuous support work and monitoring of ADTs. In particular, several meetings were held for team leaders, both in Jyväskylä and in some ADT countries.

Fortunately, given the huge amount of work facing the Project, an extension of two years was granted to the Project to enable the completion of test development, the devising of suitable feedback and advice to give to clients on completion of a test, and the construction of software for item authoring and revision, and for test piloting. In fact, even during this first phase of the Project, tests of Finnish were developed in paper-and-pencil form, piloted on approximately 400 test-takers, analysed and calibrated, and a standard-setting procedure was developed to enable test results to be related to the CEFR (see Chapter 6 and Kaftandjieva *et al.*, 1999, for details).

During the first half of 1999, the Pilot Tool and the piloting procedures were trialled in 12 countries and 23 sites, with four languages (English, Spanish, Italian and Dutch). This was done in order to test the functionality of the software to be used later for large-scale piloting and to gather feedback from pilot-site personnel and language learners on the entire assessment procedure.

By the end of Phase 1 in 1999, some 30,000 test items had been written, reviewed, revised and quality-checked and a large number input into the developing system. The DIALANG Author Tool, the Review Tool and the Pilot Tool were all complete, and the Project was ready to pilot items over the Internet in English, Dutch, Spanish and Italian.

Phase 2 of the Project, which began in December 1999, was coordinated at the Free University of Berlin, and aimed to commence Web-based piloting in all 14 languages, to calibrate items and self-assessment statements in as many languages as sufficient data could be collected for, to develop the interface of the system into a publicly available Version 1, to plan and design Version 2, which would be adaptive to the user's proficiency at the level of test items, to translate all feedback and advice into all DIALANG languages and to incorporate these into Version 1, as well as to begin to develop a Business Plan to ensure the sustainability of the Project beyond the end of public funding.

During this second phase, still funded by the Directorate General for Education and Culture, with matching contributions from the partner institutions, work progressed, albeit not without difficulty in the case of the piloting of less widely taught languages like Irish, Icelandic and even Greek. Chapter 5 describes in detail the results of the piloting for English, which was the first language to achieve sufficient numbers of candidates for calibration to take place, for the placement system to be investigated and revised and for standard-setting procedures to be refined to enable expert judgements about item levels to be gathered.

Indeed, the development of such standard-setting procedures was crucial to the Project, for two reasons. First, since the tests were based on the Common European Framework and results were reported in terms of the CEFR's six levels, it was important to be able to relate items and test scores empirically to those levels, something which cannot be achieved by the use of traditional test analysis statistics alone. Second, for those languages for which the Project was unlikely to be able to gather sufficient data in Phase 2, judgements of item difficulty and test levels were the only source of information that could be used to give users meaningful interpretations of their test performance. Chapter 6 describes these standard-setting procedures in detail and analyses their effectiveness.

The first beta version containing four languages was published in early 2002, and subjected to rigorous testing and debugging.

By the end of Phase 2 in the winter of 2002/2003, the beta form of Version 1 of DIALANG was freely available on the Internet (at www.dialang.org) in all 14 languages, with tests of Reading, Listening, (indirect) Writing, Vocabulary and Structures, all at three different levels of difficulty. Vocabulary Size Placement Tests were available in all 14 languages, self-assessment statements were also implemented in all 14 languages for Reading, Writing and Listening, and extensive feedback and advice was available for users in the system, again in all 14 languages. An alpha form of Version 2, the item-level adaptive system, was available for testing by the Project, and plans were being developed for the sustainability of the system.

At the time of writing, the DIALANG team is involved in the completion of a Business Plan, the creation of a European Economic Interest Group which could license or sell the system, and the further development of market research, marketing and publicity for DIALANG, as well as the continued piloting of test items.

Initial feedback after the launch of Version 1 demonstrated that, although DIALANG had been developed with the individual user in mind, organizations involved in language training saw its potential as a placement tool for organizing their students into homogenous groups according to language and level. What is almost ironic about this is that we seem to have come full circle. From the position we saw in Chapter 1 where it was held that placement tests were diagnostic, it now appears that diagnostic tests can be seen to be valuable for their potential as placement tests. Thus, from a position that considered diagnostic tests difficult to write, but held that any test could be used for diagnosis, including placement tests, we may now be moving to a position where tests designed to be diagnostic can also have a placement function.

Whether this is progress, only time will tell. It is clear that, like many tests, DIALANG can be used for placement – that is in a sense its macro-diagnostic function – and the fact that it claims to be able to

identify a person's CEFR level is support for this. However, much depends on whether DIALANG has sufficient quality to fulfil its diagnostic potential, and whether it can deliver feedback that is felt to be meaningful to individual learners as well as to institutions. That we will explore in the remainder of this book.

Chapter 5: Piloting in DIALANG

Test development

As described in Chapter 4, 14 Assessment Development Teams, one per language, were put together, each with a team leader, and the work of all teams was coordinated by the Coordinating Institution. The number of members of the ADTs varied from three to ten or more. All team members were experienced language teachers, usually at tertiary level, and many were already experienced test developers. Tests were developed in five major areas – Reading, Writing, Listening, Grammar and Vocabulary – at all six levels of the Framework (from A1 to C2).

In all, some 30,000 items were produced, through a process which involved introductory meetings with the ADTs to explain the CEFR, the DIALANG Assessment Framework (DAF) and the DIALANG Assessment Specifications (DAS), followed by the drafting of items, review, editing, revision and second review. The number of items produced per language varied from 3,350 (Dutch) to 525 (Icelandic). Most languages had over 2,000 items, the only other language having less than 1,000 items being Irish.

By the end of Phase 1, item pools had been developed for all 14 languages. An item pool is simply a collection of items that have been reviewed for content but which have not yet been pretested and calibrated and, in the case of DIALANG, whose layout has not yet been checked, as the preview function in the current Authoring tool cannot show the final layout of the item.

However much care is taken over test development, it is essential that all test items be piloted, to identify unforeseen problems in design, focus or wording, and to establish empirically the difficulty of each item. Thus, preparations for piloting were also begun in Phase 1.

The necessary software for piloting over the Internet was written and tested, paper-and-pencil versions of the Finnish tests were piloted on some 400 learners, and a complex design for the piloting of items in all languages was produced. In addition in Phase 1, self-assessment

statements based on those of the CEFR were written or selected, and translated into all 14 languages, and a Vocabulary Size Placement Test to precede the actual skills tests was produced for many of the languages – the remainder were completed early in Phase 2.

Piloting

In Phase 2 of the Project arrangements were made for piloting across Europe. ADTs were asked to select 60 items per skill area for piloting, following detailed guidelines produced by the Test Development Coordinators, who then reviewed the items for adherence to the DAF and DAS and to ensure consistency of layout in the Pilot Tool. Thus each item had undergone an extensive process of review and revision before piloting began.

The piloting design was initially similar for all languages: 12 pilot booklets were designed, each with 50 items – 30 of these tested one of the three major language skills and 20 were either Structure or Vocabulary items. Each test-taker only responded to one booklet, assigned randomly, which meant that any one test-taker took a test in one skill plus one in either Structure or Vocabulary. Any one item appeared in two separate booklets, to ensure greater representativeness of the pilot population. Items were selected across the range of CEFR levels (at least as far as could be ascertained by the item writer's or the ADT's judgement) which tested a variety of subskills, and which used a range of item types. Thus data were gathered on 300 items per language, representing as wide a variety as possible of the items in the item pool.

In addition, each test-taker responded to a set of self-assessment statements (in the Administration Language chosen by the test-taker) and to a Vocabulary Size Placement Test for the language in question.

The intention was to conduct an initial analysis of items, using both classical and IRT (Item Response Theory) statistics, once 100 candidates per item had been achieved (i.e. 600 test-takers in all per language since each item appeared in two of the 12 booklets). A second and definitive analysis would be conducted once 1,200 learners had taken the tests (making 200 learners per item). To date, this has only proved possible for English, the most popular language for piloting centres.

However, since all test-takers responded to the same set of self-assessment statements, and there was only one VSPT per language, there were many more responses to the self-assessment statements and the VSPT test items than to any individual test item. Since the self-assessment statements are parallel in each language, it was possible to select the final set of calibrated statements based upon the results for English, having confirmed that the calibrations were similar across languages, according to the pilot data gathered. There is thus every reason to be confident about the statistics for the self-assessments and the Vocabulary Size Placement Tests for each language.

As a result of difficulties in getting sufficient numbers of test-takers to take the pilot tests for languages other than English, the pilot design was modified, creating booklets of 60 or 70 items (instead of 50 items in the original design), but with far fewer items in the VSPT (99 instead of 150) and the self-assessment section (18 instead of 40 or so) to compensate for the greater test length. Thus only 450 test-takers would be needed before initial item calibration would be possible.

Problems encountered

The major general problems encountered in the Project were due to its complexity. Producing tests in five different skill areas at three different levels in 14 different languages is an enormous undertaking, even for an experienced and well-endowed examinations board. Across Europe, however, there is a great variety of different traditions in assessment and testing is not always a well-developed discipline. Thus the background and expertise of the ADTs varied also. This meant that the initial quality of items produced varied, and considerable effort was needed by Project coordinators to inform, train and monitor the work of the ADTs. This had particular implications for the item review process, especially for those languages which Project coordinators did not know, and thus the process was very drawn out.

The Project was very ambitious in breaking new ground, not only by applying the Common European Framework to test development, and developing diagnostic tests – an area which, as we have seen, is under-researched and under-developed in language testing – but also by developing its own software for test authoring and review, test piloting and test delivery.

It should be remembered that the tests are delivered over the Internet, for logistical reasons. In order to create item banks, it is essential to pretest and calibrate the items. Since the tests were to be delivered by computer, the Project believed strongly that it was important to pilot the items in as similar a set of circumstances as possible to the real testing situation, which meant piloting by computer and specifically over the Internet.

In fact, because the piloting software was still being developed and the Project had to deliver prototype tests to the European Commission at the end of Phase 1, as explained in Chapter 4, the Finnish items were piloted by paper-and-pencil methods. This was felt to be defensible because the majority of learners of Finnish are confined geographically, making traditional piloting more feasible. However, the same items are currently being piloted over the Internet, in order to see whether there are significant differences between the two modes of piloting.

One major problem in piloting via the Internet was with motivating centres to pilot the tests with their learners. Partly this was doubtless due to a lack of familiarity with test-taking on computer, and with inadequate, dated or insufficient numbers of computers (although the Pilot

Tool was designed to run on fairly basic systems). In part it was also due to the rather cumbersome nature of the Pilot Tool, which needs to be downloaded via the Web and then installed on each user's machine. Efforts were made to create more user-friendly instructions, and work is ongoing on creating a Web-based Pilot Tool which will dispense with the need for technical knowledge (being much more similar to programs currently running over the Web).

Network problems typical of the Web in general did not make matters easier. The lack of availability of service at the time assigned for some group sessions made it difficult to persuade the test-takers to return on a subsequent occasion, or to persuade their teachers to go through the trouble again. Some sessions had to be aborted due to local or international network demands (often on underperforming machines with too small hard disks). In at least a couple of places, it proved simply impossible to get the Pilot Tool to work at all, even though a fair amount of IT expertise was available locally.

One problem occurred because of the system design: test-takers had to complete the whole pilot test before the data was recorded into the system. With hindsight, this should certainly have been done differently because the longer the session, the greater the likelihood of a technical problem occurring, e.g. a network failure.

In addition, the tests were long: VSPT items and self-assessment statements were presented to clients even before they got to the test, and during the test there was no indication of how long the test was, how much time it might take, where they were in the test, or how many items were still to come. This may well have led some clients to give up without completing the test.

The problems were also due to pedagogical matters. First, not surprisingly, since it was not known how difficult or good the items were, test-takers could not be given any really meaningful results. They did get to know how many items they had answered correctly, but this could not be expressed in meaningful terms, e.g. in terms of the CEFR, until after the items had been piloted and calibrated. Although a rough-and-ready attempt was made to relate raw scores to CEFR levels, it seems not to have increased uptake. There was thus little incentive to students to take this test. Unlike Version 1 of DIALANG, students did not even know which items they got right and which wrong, or what the right answers were (partly because of the existence of short-answer and text-entry questions, where a range of possible answers had to be collected before the answer key could be revised).

Second, the pilot booklet design was necessarily such that nobody knew in advance which two skills they would be tested on. Therefore comparing results, even within a class, did not make much sense, since different students took different tests.

Third, the tests came 'out of the blue'. They did not relate to work a teacher may have been doing in class, and it was difficult to integrate

them into lessons after they had been administered, given the lack of informative feedback. To address this problem a 'Pedagogical Pack' was developed, which is available on the DIALANG website, which presents a series of suggestions for lessons, both preparing students to take the test by introducing the CEFR and its relevance to learning and providing a set of suggestions for follow-up activities. Although there has not been much feedback on the effectiveness of this solution, it does not seem to have made much difference in terms of uptake.

Another problem, especially given these difficulties, was finding local organizers, such as teachers, who would bring their students to the pilot test. Having enthusiastic, dedicated organizers was a key to the successes achieved in piloting, and the lack of them the reason for many a failure. Financial incentives could not be used because of limited budgets, so the reasons for some organizers being so devoted were mainly professional and personal.

And of course for some languages, there are relatively few learners anyway: Danish, Icelandic, Irish and possibly others cannot perhaps expect there to be a large population out there waiting to take a test in that language. But the number of people taking Swedish tests suggests that this may not be the only reason why some languages still have very little uptake.

Results of the piloting to date

Initially piloting was only possible for English, Spanish and Dutch, but by the end of Phase 2 in November 2002, piloting of all 14 languages was under way, and as of 14 March 2004 the number of test-takers per language was as in Table 5.1.

The pilot test population

Test-takers were asked to complete a background questionnaire before taking the pilot test proper. Although piloting is ongoing for all languages, the results for the total population included 5,154 people when these data were collated, and it is interesting to compare the total population with the sample for English, since test results for English are reported later in this volume. Data are available for all DIALANG languages, but with the exception of French, German and Spanish, they are likely to be unstable and unrepresentative. Later chapters will examine the relationship between test-taker backgrounds and test results.

The questions were all in the selected language of administration. The commonest language of administration – the language in which respondents chose to have rubrics, self-assessment, feedback, and so on – was German (25 per cent), followed by English. See Table 5.2 for percentages for all 14 languages.

Table 5.1 Tests completed per language
This information was correct at Sunday 14 March 12:53:12 UTC 2004

Test language	Number of completed tests
Danish	246
Dutch	308
English	2,265
Finnish	31
French	702
German	707
Greek	2
Icelandic	1
Irish	1
Italian	338
Norwegian	243
Portuguese	6
Spanish	680
Swedish	396
Total	**5,926**

Table 5.2 Language of administration

	Total population	English sample
Danish	7%	5%
Dutch	10%	10%
English	16%	22%
Finnish	12%	16%
French	10%	8%
German	25%	20%
Greek	1%	0.2%
Icelandic	2%	3%
Irish	0.1%	0.1%
Italian	3%	1%
Norwegian	3%	5%
Portuguese	2%	2%
Spanish	8%	5%
Swedish	4%	2%

Test-takers respond to a number of questions, in two separate sets. Once a test-taker has selected the language of administration, they are presented with the first set of questions, about their mother tongue, sex, age and education, as shown in the screenshot in Figure 5.1. Here we present only the English version.

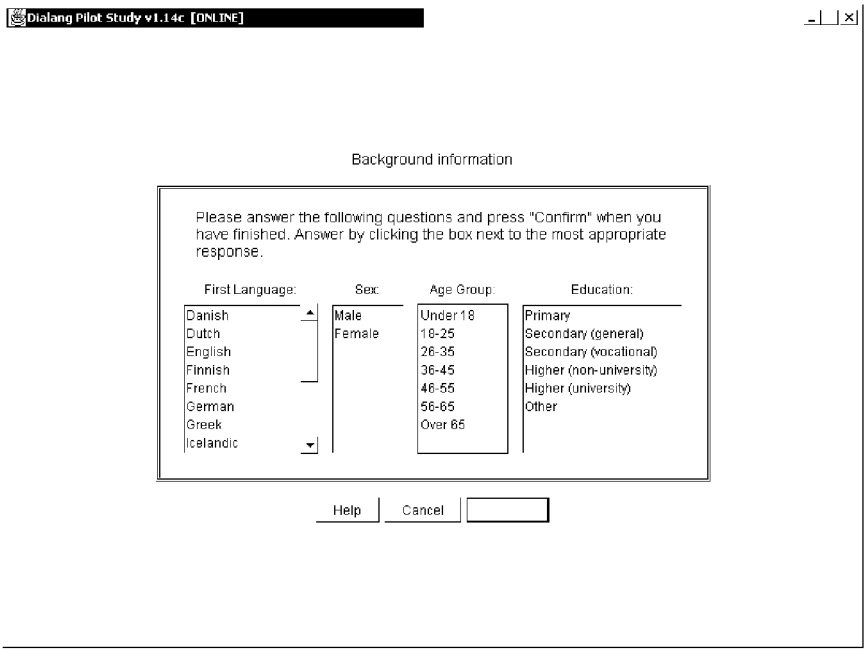

Figure 5.1 First set of background questions

The largest mother tongue group (23 per cent) are German speakers, with 'Other languages' in second place. This latter means that respondents did not have a mother tongue from among the 14 DIALANG languages. The English sample is very similar to the total population. Table 5.3 lists the figures.

Females clearly predominate in the pilot population, and the most frequent age group is between 18 and 25 – see Table 5.4 and Table 5.5.

As Table 5.6 shows, the majority of the pilot population either had or was undergoing university education, with 18 per cent still in or graduated from secondary education. The English sample has slightly fewer university students and rather more secondary school students.

Before proceeding to the next set of questions, respondents had to choose the language in which they wished to be tested. The percentage choosing each language is set out in Table 5.7. The most popular language was English, followed by French, Spanish and German.

The second set of questions relates to the length of time the respondents have studied the language of the test they have chosen and how frequently they use this language (see Figure 5.2)

There was a fairly even spread across the number of years respondents had been studying the test language, with no one group predominating (see Table 5.8). However, the English sample had a much larger

Table 5.3 Mother tongue

	Total population	English sample
Danish	6%	6%
Dutch	11%	12%
English	5%	–
Finnish	12%	16%
French	8%	8%
German	23%	23%
Greek	1%	0.5%
Icelandic	2%	3%
Irish	0.1%	0.1%
Italian	2%	1%
Norwegian	3%	5%
Other languages	18%	15%
Portuguese	2%	2%
Spanish	6%	7%
Swedish	2%	5%

Table 5.4 Sex

	Total population	English sample
Female	65%	62%
Male	35%	38%

Table 5.5 Age

	Total population	English sample
18–25	62%	63%
26–35	19%	18%
36–45	9%	10%
46–55	6%	6%
56–65	1%	1%
Over 65	0.3%	0.2%
Under 18	3%	3%

proportion of test-takers who had studied for 9 years or more and far fewer who had only studied the language for 1 to 2 years or less. This is doubtless due to the predominance of English as a foreign language across Europe.

How often respondents actually used the target language was somewhat unexpected (see Table 5.9, p. 53), with most saying 'once or twice

Table 5.6 Education

	Total population	English sample
Higher (non-university)	11%	14%
Higher (university)	59%	51%
Other	3%	2%
Primary	4%	4%
Secondary (general)	18%	20%
Secondary (vocational)	6%	8%

Table 5.7 Test language

	Total population
Danish	3%
Dutch	6%
English	42%
Finnish	0.4%
French	13%
German	11%
Icelandic	0%
Italian	6%
Portuguese	0%
Spanish	12%
Swedish	7%

a week', but a surprisingly high proportion claiming to use the language almost every day. Yet fewer claimed to speak English as frequently as they did for the other languages. The reason for this could be that many of the non-English test-takers were taking the pilot tests in the country where the language was spoken, whereas for English, more were taking the test in their own country.

Most respondents assessed their language ability as being either B1 or B2 (see Table 5.10, p. 54). Clearly these were not beginning learners, even in their own estimation. There was a relatively small proportion of very advanced learners. However, for English notably more rated themselves at B2 or above (56 per cent) than did in the total population (48 per cent). This presumably reflects the fact that most English test-takers had been learning the language for longer than those for the other languages.

Nevertheless, despite the wide spread of teaching and learning English as a foreign language in Europe, and the wide variety of languages tested, the English test population is not markedly different in its background characteristics from the total pilot test population.

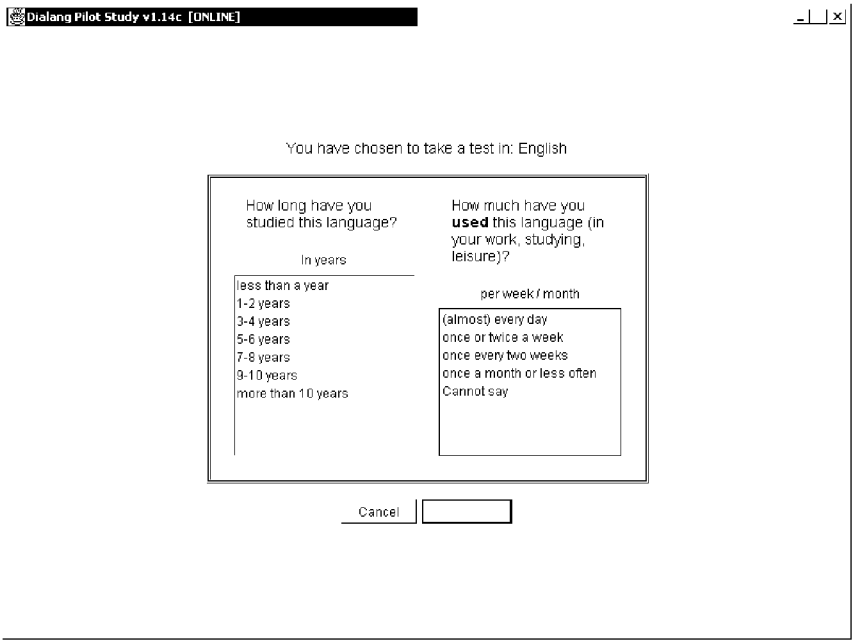

Figure 5.2 Second set of background questions

Table 5.8 How long?

	Total population	English sample
1–2 years	15%	5%
3–4 years	14%	10%
5–6 years	17%	20%
7–8 years	17%	20%
9–10 years	12%	20%
Less than a year	13%	3%
More than 10 years	14%	21%

Table 5.9 How much?

	Total population	English sample
(Almost) every day	30%	25%
Cannot say	9%	9%
Once a month or less often	12%	13%
Once every two weeks	6%	8%
Once or twice a week	43%	44%

Table 5.10 Level of self-assessment

	Total population	English sample
A1	10%	8%
A2	13%	10%
B1	29%	25%
B2	27%	29%
C1	12%	16%
C2	9%	11%

Comparison of pilot and Version 1 test items for English

We now turn to the results of piloting for English, the only language for which meaningful data is currently available. Although the data initially used for calibration purposes was based on 1,200 cases (November 2001), subsequent analyses were made of a larger data set gathered in the interim (2,065 cases). It is of interest to note that as a result of the recalibration, only one further item, a Grammar item, had to be dropped from Version 1. Nevertheless, the results that follow are based on the recalibrated data.

The results of piloting were very encouraging, within the limitations of the booklet design used. As Table 5.11 shows, less than 10 per cent of the items (24 out of 300) were lost – that is, surprisingly few items failed to calibrate using IRT. That may in part be due to the fact that the pilot population was fairly heterogeneous, but it is probably also due to the quality of the items themselves and to the careful and extensive review process undertaken before the booklets were assembled. Relatively more Reading items failed to calibrate, whereas all Vocabulary items survived.

Several of the Reading items that were rejected were open-ended – short-answer or text-entry – for which there were too many acceptable responses. Two items were at level C2, and there was no clearly acceptable answer. In terms of subskills, four items testing 'reading intensively

Table 5.11 Number of items after calibration

	Frequency	Percent of total	Percent surviving
Grammar	56	20%	93%
Listening	56	20%	93%
Reading	50	18%	83%
Vocabulary	60	22%	100%
Writing	54	19%	90%
Total	276	100	92%

Table 5.12 Number of items by subskill (calibrated items)

Skill	Subskill	Number of items	%
Grammar	Morphology – Adjectives and Adverbs – comparison	4	7.0
	Morphology – Adjectives and Adverbs – inflection	1	1.8
	Morphology – Nouns – definite/ indefinite articles	2	3.5
	Morphology – Nouns – proper/common	4	7.0
	Morphology – Numerals inflection	8	14.0
	Morphology Others	6	10.5
	Morphology – Pronouns – context	6	10.5
	Morphology – Verbs – active/passive voice	4	7.0
	Morphology – Verbs – inflection, tense, mood, person	6	10.5
	Syntax – Organization/realization of parts of speech – word order statements, questions	10	17.5
	Syntax Punctuation	4	7.0
	Syntax – Simple sentences vs complex sentences – coordination	1	1.8
	Total	56	98.1
Listening	Identifying Main Idea/Distinguishing from suppporting detail	34	60.7
	Inferencing (including lexical inferencing)	16	28.6
	Listening intensively for specific detail	6	10.7
	Total	56	100.0
Reading	Inferencing (including lexical inferencing)	25	50.0
	Identifying Main Idea/Distinguishing from supporting detail	17	34.0
	Reading intensively for specific detail	8	16.0
	Total	50	100.0
Vocab	Combination	13	21.7
	Meaning	19	31.7
	Semantic Relations	16	26.7
	Word Formation	12	20.0
	Total	60	100.1
Writing	Knowledge of accuracy (grammar/ vocab/spelling)	22	40.7
	Knowledge of register/appropriacy	15	27.8
	Knowledge of textual organization (Cohesion/coherence, text grammar/ paragraph organization)	17	31.5
	Total	54	100.0

for specific detail' were rejected (out of 12 piloted), five items testing 'inferencing' were rejected (out of 30 piloted). Only one item testing 'identifying main idea' was rejected (out of 18 piloted).

In Grammar, four items were rejected, one testing the 'comparison of adjectives and adverbs' (out of five piloted), one testing 'morphology' and two testing 'syntax – punctuation'. More items were rejected in Writing: all six rejected items were testing 'knowledge of accuracy' (out of 28 piloted). In this latter case, there were doubts about what was being tested, and in two cases there were too many acceptable answers.

Table 5.12 shows the breakdown of subskills across the various skill tests, for surviving items.

In terms of item types, the majority of pilot items were traditional multiple-choice (153 items out of 300), and there were very few short-answer items (10 out of 300). However, this varied by booklet and skill, with Writing booklets having far fewer multiple-choice items and relatively more drop-down items. Multiple-choice items were especially preponderant in Listening tests, with remarkably few short-answer items.

There was also a clear preference for text-entry items for Vocabulary and Writing, the aim being to measure active knowledge rather than passive knowledge of Vocabulary and for less indirect tests of writing to be included where possible.

Table 5.13 shows the distribution of item types by test skill, for those items that survived calibration. Relatively few short-answer items survived calibration (three out of ten piloted) but nine multiple-choice items (out of 153 piloted) also failed to calibrate, and five (out of 74) text-entry items. Three (out of 63) drop-down items failed to calibrate.

Table 5.14 gives details of how many items were judged by their authors to be at each level of the CEFR (ABG = Author's Best Guess), and how many items survived calibration. The most attrition was suffered by items at A2, yet no items at A1 were rejected. Otherwise there was a fairly even pattern across the levels of items not calibrating. It is interesting to note that there were relatively few C2 and A1 items, even though item writers had been asked to produce roughly equal numbers

Table 5.13 Item types by skills, calibrated items only

	Drop-down	Multiple-choice	Text-entry	Short-answer	Total
Grammar	9	32	15	–	56
Listening	–	55	–	1	56
Reading	10	39	1	–	50
Vocabulary	19	12	27	2	60
Writing	22	6	26	–	54
Total	60	144	69	3	276

Table 5.14 Number of items at CEFR (ABG) levels

	Piloted items	Calibrated items	Percentage surviving
A1	31	31	100%
A2	66	56	85%
B1	76	73	96%
B2	53	50	94%
C1	53	48	91%
C2	21	18	86%
Total	300	276	92%

at each level. It also appears that A2 and B1 were the most frequent levels of items (at least according to the Author's Best Guess). This may reflect a perception that most language learners taking these tests would be either A2 or B1. It may also be the case that it was easier to write items at these levels (although in the event, more A2 items were lost than any other level), or it may reflect the fact that most item writers were tertiary level teachers rather than secondary teachers.

In terms of estimated CEFR levels, the distribution by skill is given in Table 5.15, which shows the number of items in the pilot, and in Table 5.16, which shows the number of items that survived calibration. The levels were unevenly distributed across the different skills in both the pilot and Version 1, with Vocabulary showing the most even

Table 5.15 Number of items, by skill and estimated level (ABG), in the pilot

Skill/Level	A1	A2	B1	B2	C1	C2
Grammar	8	11	14	13	10	4
Listening	4	18	23	9	5	1
Reading	6	13	12	12	11	6
Vocab	9	12	12	8	12	7
Writing	4	12	15	11	15	3

Table 5.16 Number of items, by skill and estimated level (ABG), surviving the pilot

Skill/Level	A1	A2	B1	B2	C1	C2
Grammar	8	9	13	12	10	4
Listening	4	16	22	9	4	1
Reading	6	7	11	12	10	4
Vocab	9	12	12	8	12	7
Writing	4	12	15	9	12	2

Table 5.17 Mean scores (in percentages) for each skill and language element, for 12 booklets (b25–b36)

	b25	b26	b27	b28	b29	b30	b31	b32	b33	b34	b35	b36
Total score	76	70	69	76	60	66	64	61	76	73	67	71
Reading	68	66	65	71								
Grammar	87	76	76	84	76	81						
Writing					49	56	61	56				
Vocabulary							68	69	77	73	69	70
Listening									75	73	67	71

distribution, and Listening the least. It was Reading at A2 (ABG) that suffered the highest number of rejected items.

Of course, what is more important than the Author's Best Guess is how the items are distributed in terms of actual difficulty. Table 5.17 presents the mean difficulties, in percentages, of each skill, by booklet.

Remembering that different clients took each booklet, and that each item appeared in only two of the booklets, these figures must be interpreted with caution, since variation in means can be due to different abilities in people or different difficulties in items.

Clearly the variation in total score across booklets is caused by the greater difficulty in general of the Writing tests, and the greater ease, in general, of the Grammar tests. However, the difficulty of each item was calculated on the basis of Item Response Theory (using the program OPLM), which allows one to calculate item difficulty independently of the ability of the test-takers attempting that item. Thus we

Table 5.18 Cronbach alpha reliability of each test booklet

Booklet number/ content	N taking the test	Alpha Total	Alpha Skill (k = 30)	Alpha Language (k = 20)
1: Reading/Grammar	182	.861	.817	.70
2: Reading/Grammar	171	.881	.821	.750
3: Reading/Grammar	178	.914	.859	.843
4: Reading/Grammar	187	.889	.833	.784
5: Writing/Grammar	182	.896	.856	.746
6: Writing/Grammar	184	.899	.865	.743
7: Writing/Vocabulary	183	.929	.875	.862
8: Writing/Vocabulary	185	.923	.881	.835
9: Listening/Vocabulary	162	.911	.887	.789
10: Listening/Vocabulary	168	.917	.900	.792
11: Listening/Vocabulary	135	.917	.884	.828
12: Listening/Vocabulary	141	.904	.865	.825

can be confident that the calibrated difficulty of each item that survived is unaffected by the ability of the pilot population.

Finally, the reliabilities for each booklet were entirely satisfactory, as Table 5.18 shows.

Lessons learned

1 The decision to go for computer-based, Internet-delivered testing has been described earlier, and it was both sensible and practical. It did, however, mean that a team of programmers, system developers and managers had to be engaged. The Project had to decide how to test what had to be tested, design templates that item writers could use that would deliver user-friendly test items, and develop the programming infrastructure for a complex project which involved item authoring and review, piloting, further reviews and revisions and the capture, analysis and storage of data, as well as storage of the items themselves. This inevitably made piloting more complex than normal paper-and-pencil piloting.

2 The Project required the tests to be Web-delivered, relatively secure, interactive and responsive. At the time (1997) the only technology that met those requirements was Java, and so a decision was made to program in Java as that seemed to be the most future-proof. Technology moves so fast that now there are a number of methods which meet such requirements. With hindsight, an XML-based Web-browser-borne system would have been preferable, but even that conclusion may look wrong in a couple of years' time. Part of the problem of ensuring adequate piloting was that the pilot program was being developed and debugged at the same time as the program was being used, which caused much frustration on all sides.

3 For political reasons it was felt important that as many of the originally proposed 15 languages as possible should be involved from the very beginning. This caused enormous problems of team creation, liaison with potential partner institutions, and cross-Project coordination and management. It made life hugely more complicated and expensive, and frankly offered little in the way of advantage other than the political one of being able to say that the less widely taught languages were clearly involved from the very beginning. There would, of course, have been no objection to one such language being involved from the beginning: the issue is not which language, but how many.

4 To summarize the problems faced in piloting described earlier, these included:
   i) inadequate equipment, network connections or available local expertise to handle technical problems;
   ii) a relatively unfriendly Pilot Tool which only gathered data at the end of the session;

iii) poor network connections leading to data being lost/students frustrated;
iv) long and tiring tests, including a long VSPT and Self-Assessments, with no indication of test length or number of items remaining;
v) relatively sparse feedback on performance, although numerical results were available immediately;
vi) little incentive for students to take a test, or for administrators to persuade teachers or students;
vii) no immediate connection of piloting with local pedagogy/syllabus;
viii) no student knowing in advance what sort of test (which skill) they will be taking and results not being easily comparable;
ix) lack of dedicated enthusiastic local organizers of piloting;
x) few learners for some languages;
xi) inadequate financing and resourcing.

5 Despite all these problems, we are confident that the data gathered during piloting and reported above attests to the quality of DIALANG items. Users of Version 1 are, where appropriate, presented with a screen which cautions them that the test results may not be as stable as might be desired. Yet this does not appear to have dissuaded users from entering the system and taking the tests.

In later chapters, I describe the results for each test, skill by skill. But first, in Chapter 6 I will describe how the test items were related empirically to the Common European Framework, since a major diagnostic function of DIALANG is to inform learners what level they have reached in the skill or language aspect they selected, in terms of the six-level Framework.

Chapter 6: Setting standards

Why standard setting is necessary

'*The aim of any standard setting procedure is to establish a cut-off point between different levels of mastery, based on which the decisions can be made about personnel classification, selection, certification or diagnosis*' (Kaftandjieva *et al.* 1999).

It is important to know what any performance on a test actually means, which implies that we need to be able to interpret a test score. How is one to decide whether a test-taker has reached the standard required for a 'Pass' or whether the performance is such as not to merit being considered adequate, and thus is a 'Fail'? The answer is that we need procedures to determine what the standards are that have to be met, and these procedures are known as standard-setting procedures. In any test, it is important to decide what level a candidate has to be at in order to be considered to master the content or skills being tested.

One way of doing this is simply to compare a candidate's score with the scores of a larger population who have already taken the test. One can then decide that if a test-taker has reached a score which is higher than a target percentage (say 80 per cent, or 60 per cent) of the population – called a norming population because that population sets the norm for the test – then the test-taker can be considered to have been successful. The percentage of the population whose performance has to be matched or surpassed will depend upon the purpose of the test, or on issues like the availability of places for which the test is being taken, and the expectations of test users as to what proportion of the population 'deserve' to 'pass'. This is known as norm-referencing, where norms are set by simple comparisons with some norming or target group, regardless of the actual, 'absolute', level of achievement. Clearly, the greater the difference between the characteristics of the norming population and the sample of candidates taking the test, the harder it will be to interpret the test scores.

Alternatively, in many settings an arbitrary pass mark is used, for example 50 per cent of the items correctly answered or 75 per cent of

the items or 40 per cent, whose justification may well be lost in the mists of time. The problem with this method of setting pass marks or defining standards is that the standards will vary depending upon the difficulty of the test and the test items. Forty per cent correct of a set of very difficult items is a better performance than 40 per cent correct of a set of easy items, and probably better than even 80 per cent correct of the easy items. Thus the meaning and fairness of such arbitrary pass marks will depend upon the actual difficulty of the test being taken (and in many settings, the difficulty of the test varies from year to year and is not established empirically in advance by pre-testing).

A further alternative might be to compare performance on the test in question with candidates' performances on another, highly similar test, whose results are known to reflect the standards required. In most cases, however, no such alternative tests exist which can be said to measure the same abilities (and if they did, why would one require candidates to take a different test?).

In many settings, there may be no historical pass mark, no population on whom the test can be normed, no desire to decide on the value of a candidate's performance by comparing it with the performances of other people, whose abilities may in any case be unknown, and no other suitable test which can be used for purposes of comparison – in short, no norms.

DIALANG and the problem of standards

In the case of DIALANG, there were no such norms, or norming population. Indeed, by the very nature of the Project, whose tests are freely available on the Internet to anybody who cares to take them, it is practically impossible to define the population for whom the test is intended. Nor are there any alternative tests with which the DIALANG test results can be compared, and there is no way of knowing what proportion of items a person needs to get in order to be considered to have achieved a given CEFR level in a particular language or skill. For example, nobody knows what proportion of the potential test population is capable of getting a B1 in Reading, or an A2 in Listening. Since the difficulty of a test will vary according to the test items it contains, there is no obvious *a priori* way of saying that in order for a person to be considered to be a B2 in Writing, for example, that person must achieve 50 per cent correct (or any other figure plucked out of the air).

What is needed is some means of interpreting a test score independently of the performances of the people who happen to take that test. What we need in short is some way of defining what are technically called mastery levels. We need to be able to answer the question: at what point can a person's performance on this test be considered to reflect a mastery of the ability being tested, at the level required?

DIALANG faces much thornier problems in setting standards than those faced by many other test developers who develop tests for clearly defined, monolingual populations, simply because of the multilingual nature of DIALANG's target audience. It is clearly the case that learners from one first language (L1) background will find some aspects of a given foreign language more difficult (or easier) than those coming from a different L1 background. It is believed that although the difficulty of reading comprehension items is more stable, no matter which L1 background test-takers come from, grammar and vocabulary test items are less stable, reflecting the closeness of L1 and the target language.

The implications for systems like DIALANG are, firstly, that it is more difficult to establish a stable standard for each test and test language across all languages in the DIALANG target population, and, secondly, that the standard-setting procedures must be shown to work well for different L1 backgrounds. This is an issue that remains to be studied in detail, but which is nonetheless important for all that.

Using the CEFR in setting standards for DIALANG

As we briefly discussed in Chapter 4, standard setting was a very important component of the process of developing, piloting and calibrating the DIALANG test items. DIALANG items had been developed according to the DIALANG Assessment Framework (DAF) and the DIALANG Assessment Specifications (DAS), which were themselves based upon the Common European Framework (CEFR). Test results were to be reported according to the six main levels of the CEFR, and raw scores, or any form of figures, would not be reported (other than for the Vocabulary Size Placement Test, itself not related to the CEFR), on the grounds that numerical results mean very little or nothing to users without a frame of reference to help them interpret the results.

Although the DIALANG test items were based upon the CEFR and results would be reported according to the CEFR, it was still necessary to develop a means of relating both test items and resulting scores to the CEFR. Although test items were developed by item writers according to the CEFR as far as possible, the CEFR itself in its present form is not always sufficiently explicit about what exactly a test or a test item at a given level of the CEFR should actually contain. Therefore, even with the best of intentions, item writers might well have produced items which, although intended to be at a given level of the CEFR, might actually have been a level or more higher, or a level or more lower than the one intended. This is, of course, precisely the reason why test items have to be piloted, in order to see whether the predicted level of difficulty matches the actual level of difficulty experienced by test-takers.

In order to know whether a given item is indeed at the level of difficulty intended, therefore, pre-testing/piloting on suitable samples of

test-takers is crucial, as described in Chapter 5. The results of such piloting are a central element in deciding whether an item meets the standard intended.

However, pre-testing and calibrating items onto a common scale of difficulty is not enough, since it does not provide guidance as to where on the scale a person can be said to be A2 rather than A1 or C1 rather than B2. What is needed, in short, is some way of deciding where the cut-off point comes on the scale of difficulty between one CEFR level and another. One way to do this is to use the judgements of experts, people who are very familiar with the Council of Europe Framework, who know exactly what is meant by each level of the Framework.

Two approaches to standard setting

Two alternative sorts of judgement can be gathered: one centred on candidates, who then take the test(s), and one centring on the items that the test contains. The first method, known as a person-centred standard-setting procedure, requires judges to know a set of candidates very well in terms of their language ability. They are asked to assess the ability of each candidate according to the level descriptors of the CEFR, with which they must also be very familiar. Each candidate thereby receives a CEFR level according to the judgement of the expert – who may be a teacher or some other professionally qualified person. Each candidate then takes the test in question, and their performance on the test is compared with the previous judgement about their ability on the CEFR scale. By looking at a sufficient number of candidates and comparing their scores and their CEFR levels as judged by experts, it is possible to reach an estimate of the point on the scale of scores where a person can be said to be at a given CEFR level rather than one below or above. In other words, cut-scores can be defined in this way.

In many circumstances, however, such procedures are impractical. Certainly in the case of DIALANG, which is intended to be used across Europe and even beyond, it would be very difficult if not impossible to gather an adequate sample of the range of people who might take a DIALANG test and to get reliable and comparable judgements about their abilities, and then to arrange for those same people to take the same DIALANG test in order for their scores to be compared to their predetermined CEFR levels.

Instead, a second procedure is more practical, which involves making judgements, not about candidates, but about test items. This procedure, a test-centred procedure, also requires expert judges who are thoroughly familiar with the CEFR for the particular skill being judged/tested. Essentially the judges' task is to look at each item that the test contains and to decide for each item whether it can be mastered (answered correctly) by a person at a given level of the CEFR. If enough expert judges are available, their judgements can be

compared with each other statistically. Depending upon the specific procedures used to gather the judgements, the analyst can spot a judge who may be out of line with the majority, or who may be erratic, sometimes judging an item to be at a low level, for example, other times judging the same item to be at a higher level. Such inconsistent judges can be removed from the panel of experts, and only those judges will be used who have been shown to be consistent and in agreement with the other judges. Once a panel of reliable, consistent judges has been gathered, then their judgements of the level of any item can be aggregated to give the definitive CEFR level for that item.

These items can then be placed onto a scale, such that items judged to be easy appear lower than items judged to be harder. The empirically established difficulties of the same items are then compared against the judged CEFR levels of the items, and the point on the scale of difficulty at which items change from being at one level to a level higher can in principle be taken to be the cut-score between the two CEFR levels.

At this point it is important to point out that, through the application of Item Response Theory, it is possible (and crucial in standard setting) to place the difficulty of test items and the ability of people being tested onto a common scale (known technically as a logit scale) which matches people's ability to the difficulty of items. Through IRT the difficulty of items can be estimated independently of the ability of the people who answered the items and the ability of people can be estimated, on the same scale, independently of the difficulty of the items. This is not possible in classical item analysis since the difficulty of an item will necessarily reflect the ability of the people taking the item: an item taken by able people will appear easy, whereas the same item taken by less able people will appear more difficult.

Standard-setting procedures used in DIALANG

In this chapter, we will examine the different procedures developed by the DIALANG Project in order to be able to set cut-scores for the many different tests in the system, so as to be able to interpret any given candidate's score in terms of the CEFR. It is only through the use of appropriate standard-setting procedures that we are able to say that a person with a given test score is at level X on the CEFR. Without standard setting, we would merely have a numerical score for a person but no way of interpreting that score in terms of the CEFR. Thus, for the diagnosis of a person's ability level in terms of the CEFR, standard setting is absolutely crucial.

In fact, during the DIALANG project, two different standard-setting procedures were developed and experimented with, in order to find a method which was practical in terms of time required, accurate in terms of reliability, and which yielded meaningful results.

The quality of judges

However, of more importance than the differences among the procedures used, which will be detailed shortly, is the quality of the experts involved. No procedure, however sophisticated, can compensate for using judges who do not have the necessary expertise and experience, and the most important factor in any standard-setting procedure is the quality of the judges. In the case of DIALANG, therefore, it was of the utmost importance that we identified experts in the language being examined, with sufficient expertise in and knowledge of the skill being tested, preferably with relevant experience in both language testing and language teaching, and also with a higher degree in applied linguistics.

In addition, however, it was also important that the experts should have a good working knowledge of the CEFR, since it was according to the levels of the CEFR that the judges were going to pass their judgements. In fact, this was easier said than done, because the CEFR itself was relatively new, and not necessarily widely known across the different assessment and linguistic cultures represented by the DIALANG languages. Initially, after all, the CEFR was only available in English, then later in French and only recently has it been translated into other European languages. Moreover, the CEFR changed somewhat as reactions to it were received during piloting, and so the published 2001 version is somewhat different from the 1996 version available at the beginning of the DIALANG Project.

An important matter to consider, therefore, is how well the experts actually know the CEFR, however expert they may be in general terms about language, language education and language assessment. Experience shows that many people, including so-called experts, claim a knowledge of the CEFR, where in reality they are only familiar with part of it. In fact, the CEFR is a document seeking to describe a highly complex activity, that of language use and language learning, and so it is inevitably something of an encyclopaedia about a very large and broad topic, and even working with the CEFR for a number of years does not guarantee familiarity with every aspect of it.

Therefore, whilst seeking to identify expert judges who claimed considerable familiarity with the CEFR, who also had suitable expertise and experience, it was important to ensure a roughly equal familiarity with the CEFR before judgement sessions proper began.

Familiarization procedures

First, a group of experts, which would ideally consist of about a dozen people with suitable backgrounds, was assembled for each language. Between 7 and 12 judges are recommended for each DIALANG procedure (although, in practice, DIALANG had to be content with only five judges in the case of some languages).

Judgements were made skill by skill, such that, first, all necessary judgements were made about, say, Reading before another skill was judged. Judgement sessions might be a few hours apart, or as much as a week apart, depending upon the procedure used and the availability of judges. At the beginning of each judgement session, however, it was essential to hold one or two familiarization sessions. This involved, first, outlining the CEFR and its philosophy in general and discussing relevant aspects of that in relation to the skill being judged. Then judges were given a sorting task which involved them in putting individual Can-Do statements for the skill in question into the correct CEFR level. This was done individually, then results were pooled and disagreements among judges as well as with the original CEFR levels were discussed until agreement was reached. It was important that judges accept the CEFR levels and did not insist on their own interpretation, since the whole point of the exercise was to reach a common understanding of the CEFR and then apply that to test items.

Once agreement had been reached, a second sorting exercise took place using the DIALANG descriptors for the skill in question and again discussions of disagreements were held until agreement was reached. In every case, the discussion of DIALANG descriptors reached agreement much quicker than the initial discussion of the CEFR descriptors, thereby suggesting the effectiveness of the initial sorting and discussion activity. Only once agreement had been reached did the judgement sessions proper begin, and it is important to stress that this familiarization phase is of crucial importance to the success of the remaining phases.

The judgement procedures used in DIALANG

In DIALANG two different procedures were developed in order to gather judgements. Each procedure was preceded by a familiarization session along the lines just described.

Each procedure involved making iterative judgements about individual test items. In essence, judges make a decision about whether a learner at a level defined according to the CEFR level descriptors can complete an item successfully. They repeat this process for each item, for all five skills. The details of how this is carried out vary from procedure to procedure, as described below.

The standard-setting stage proper is conducted by a data analyst once the empirical calibration results from the piloting of the language concerned are available. This involves running a program specially developed for the Project on the data (the calibration values and the expert ratings).

Standard Setting (SS) (Version 1) procedures

During DIALANG Phase 1 standard-setting procedures were developed and applied to Finnish test items. In Phase 2 these were also

applied to English and Spanish items. The procedures are described in the Standard Setting Manual for Phase 1 (Kaftandjieva *et al.*, 1999) and in papers and articles (Kaftandjieva and Takala, 2002; Kaftandjieva, 2004).

First, there is a brief training session, in which the judges are provided with six items which are not included in the Pilot Study and which, according to their item labels, belong to different levels of difficulty. The judges are asked to examine the first item in the list and to answer the following question: '*Do you agree (yes/no) that a person with language proficiency at level B2 should be able to answer the item correctly?*'

Answers are collated, the percentage of agreement is computed and the possible reasons for disagreement are discussed until consensus is reached. The training procedure continues in the same way with the remaining five items.

The actual judgements begin with the first CEFR level to be judged (which need not be A1). The judges are given a booklet containing the 60 pilot items (with the key) for a given skill. They are required to answer the question '*Do you agree (yes/no) that a person with language proficiency at level Z should be able to answer the following item correctly?*' for each of the items, filling a judgement form *Z* (*Z* = A1, A2, B1, B2, C1, C2). They go through the whole set of items for a skill, deciding for each item whether the person can get it right or not. The data are simply Yes/No judgements. When they have completed this for one level, a different level is taken, and they make their judgements again. The six levels are presented to the judges in random order. During each of the steps the judges have no access to the forms completed previously, and each judge proceeds independently of the others. The task assessments are skill-specific, and one skill is covered per day, except possibly for vocabulary and grammar, both of which might be rated on one day, since the items are typically shorter and more focused, and thus quicker to process.

The procedure described above results in six judged levels for each item. A 'Yes' response is represented with the figure 1, and so the pattern of responses 1 1 1 1 1 1 means that the judge thinks the item is an A1 item, because s/he has responded 'Yes' to each of the six levels. The pattern 0 1 1 1 1 1 means that the item cannot be answered correctly by a person who is at A1, but it can be answered correctly by anybody who is at A2, or any level above A2. 0 0 0 0 0 1 means C2, and all zeros would mean it is above C2 which would be coded as 7.

Patterns like 0 1 0 1 1 1 or 1 0 0 1 1 1 clearly show inconsistencies in judgements and need to be interpreted. Such patterns are a minority but they do occur, in varying proportions depending on the judges.

One way to decide on what the inconsistent response patterns mean is to substitute the inconsistency with the mode of the other judges' judgements.

A different method could be used for cases where the judge seems to hesitate between (at least) two levels, as shown by a pattern where there

is one zero following the first 1, e.g. 0 1 0 1 1 1. In this example, one could assume that the judge was hesitating (at least) between levels A2 and B1. In such cases, one could

- use the lowest level where there is 1, i.e. A2 in this case
- use the higher level, i.e. B1 (the level at which the 0 is located) or even
- use the level where a consistent pattern of 1s starts, i.e. B2 here.

Deleting the judge with many inconsistent responses is an obvious solution. However, almost everybody showed some inconsistency and so decisions had to be made individually in some cases. Furthermore, many teams were rather small to start with, so deleting even one or two judges from a team of five would cause a significant reduction in judgements.

Difficulty Estimation Standard Setting (DESS) (Version 2a) procedures

SS Version 1 was only applied to Finnish in Phase 1 and then English and Spanish in Phase 2, as noted above. In DIALANG, it was always going to be difficult to get sufficient numbers of candidates in piloting to be able to calibrate the items for certain less widely taught languages (Irish, Icelandic and possibly Greek and Portuguese). And so, in the early part of Phase 2, a procedure was developed called Difficulty Estimation, in the hope that this might overcome the need for piloting and calibration of items, or at least that it might provide a useful intermediate stage until such time as sufficient numbers could be achieved in piloting. A method was devised whereby judges were to examine tasks, item by item, and make a judgement as to each item's difficulty. This they did by estimating the percentage of items that a learner at a given level would get correct, if they were to take 100 similar items. They were also asked to estimate the minimum and maximum likely percentage. The hope was that a suitable design of rotating booklets, together with a sufficient number of language experts, would result in reasonably robust estimates of difficulty, at least until a suitable number of pilot test-takers had been achieved.

The details of the procedure are, however, irrelevant to this discussion because correlations between estimated difficulty and actual empirical difficulty for both English and Dutch proved to be no greater than .5, accounting for only a quarter of common variance. Such low predictions of difficulty are not acceptable and the Difficulty Estimation procedure had to be abandoned.

However, in addition to the difficulty estimation procedures (the DE in the acronym DESS), the Project experimented with a second procedure for standard setting, known as a 'piling' procedure (the SS in the acronym DESS).

This task, called 'Standard Setting', involved each judge sorting all 60 items for any one skill into one of seven piles (A1 to C2+) and then ranking each item for difficulty within each pile. The task instructions were to start the construction of piles with an A1 pile. For this pile, judges were to choose all the items that a learner at level A1 should **master**. For each item, they were asked to answer the question: 'Is it reasonable to **require** that a learner at level A1 gets this item right?' If the answer was 'Yes', the item belonged to the A1 pile. They were then instructed to construct the A2 pile, and for each remaining item to answer the question: 'Is it reasonable to **require** that a learner at level A2 gets this item right?' If the answer was 'Yes', the item belonged to the A2 pile. It was pointed out that in this way they were also implying that all the items that they put in the A1 pile could be mastered by a learner at level A2. This procedure was to be continued through the remaining four levels. Any items remaining after the C2 pile had been constructed were to be labelled C2+.

Once judges were relatively sure that each item was in the right pile, they were asked to order the items WITHIN each pile, according to their difficulty. The resulting data were then analysed by the Project.

Standard Setting (STSE) (Version 2b)

As a result of the disappointing results with the DESS procedures, which combined both standard setting and difficulty estimation, a revised standard-setting procedure was developed, as follows.

Five to ten expert judges were selected per language, who were already familiar with the language being judged and with the CEFR and DIALANG scales, and who had relevant teaching or testing experience. The judgement session consisted of three phases: a (re)familiarization with the CEFR and DIALANG scales; a first round of standard setting; a second round of standard setting completed by the same judges on the same skill to enable an estimate of the reliability of judgements.

The session was supervised by a leader (unlike the SS part of the DESS, which was carried out at home after the Difficulty Estimation procedures had been completed in a group session). The familiarization phase involved, as usual, a sorting task and a plenary discussion of the results, as described in SS Version 1.

In the first round of standard setting, each judge, working independently, sorted all 60 items within a skill into seven piles (A1 to C2+), as described above in the DESS. Once they had allocated all the items to one pile or another, they were asked to go through each pile once more, to check that they were satisfied that a particular item belonged to the pile where it had been placed – or whether it should go to a pile lower or higher. They were not asked to rank order the items within a pile in order of difficulty (unlike the DESS procedure).

After an interval of 2–3 hours, each judge then completed a second sorting of all 60 items within a skill into seven piles, without reference to their earlier sorting.

Comparison of standard-setting procedures

It can be seen that although these procedures have elements in common, they vary in complexity and length of time required for completion. Version 2a (DESS) was intended to kill two birds with one stone, in the hope that estimations of difficulty might replace actual empirically derived difficulties. The standard-setting component was rather short. The Version 2b (STSE) is considerably simpler and less time-consuming than Version 1 (SS) and also has the advantage that it allows a direct check on rater consistency which is much less complex than the check described above in Version 1.

Table 6.1 summarizes the essential differences of each standard-setting procedure. In each case, there was a preliminary familiarization stage, as described in SS Version 1 above.

The different standard-setting procedures were applied to the 14 DIALANG languages as follows:

- SS Version 1: Finnish, English, Spanish
- DESS Version 2a: English, Italian, Dutch

Table 6.1 Essential differences between the judgement procedures

SS Version 1	**DESS Version 2a**	**STSE Version 2b**
One day per skill	Two sessions per skill	All skills in 1–2 days
1) *Judgement training*: 6 items 2) *Judgement session*: For each item out of 60, 6 separate judgements: '*Do you agree (yes/no) that a person with language proficiency at LEVEL B2 in* ***English*** *should be able to answer the following items correctly?*' Thus each item judged 6 times.	1) *Standard Setting*: Each judge sorts 60 items into one of seven piles (A1 to C2+). For A1 pile '*Is it reasonable to* ***require*** *that a learner at level A1 gets this item right?*' 2) Judge then ranks each item within each pile in order of difficulty. Each item judged and ranked once only.	1) *First round*: Each judge sorts all 60 items for any one skill into one of seven piles (A1 to C2+). For each item: '*Is it reasonable to* ***require*** *that a learner at level A1 gets this item right?*' 2) *Second round*: Process repeated after 2–3 hours.

- STSE Version 2b: French, German, Danish, Greek, Icelandic, Irish, Norwegian, Portuguese, Swedish

The results of the different procedures are quite complex. However, in the final section of this chapter, I show some of the main findings.

Results

Intra-rater reliability or consistency

As explained above, the SS Version 1 allowed a detailed examination of the inconsistencies of judgement of items. The number of inconsistently judged items can be considered as an indication of intra-rater reliability since the judges judged the same items several times.

As an example Table 6.2 displays the number of inconsistently judged items for each of the 12 judges of Spanish, skill by skill:

- small inconsistencies (101111, 010111, 001011, 000101, 000010)
- greater inconsistencies (100111, 100011, 010011, etc.)
- the total number of inconsistencies (the sum of the small and greater inconsistencies).

Table 6.2 Rater inconsistencies across items and skills

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10	Rater 11	Rater 12
Listening												
– small	0	0	0	0	0	0	2	0	0	0	0	7
– big	0	0	0	0	1	1	1	0	0	0	0	16
– total	0	0	0	0	1	1	3	0	0	0	0	23
Reading												
– small	0	0	2	9	2	1	0	2	0	n/a	0	
– big	0	0	1	0	1	0	0	1	0	n/a	0	
– total	0	0	3	9	3	1	0	3	0	n/a	0	
Writing												
– small	2	0	0	3	0	0	2	0	0	7	0	
– big	0	0	0	1	0	0	0	0	0	0	0	
– total	2	0	0	4	0	0	2	0	0	7	0	
Vocab												
– small	0	0	6	2	1	0	1	1	1	0	0	0
– big	1	0	3	0	0	0	0	0	0	0	2	0
– total	1	0	9	2	1	0	1	1	1	0	2	0
Structures												
– small	0	3	0	0	3	2	0	1	0	2	1	0
– big	0	0	2	1	0	0	0	0	0	0	34	3
– total	0	3	2	1	3	2	0	1	0	2	35	3
Grand total	3	3	14	16	8	4	6	5	1	9	37	26

For Listening, Judge 12 was by far the most inconsistent rater. For Reading, only Judge 4 was somewhat more inconsistent than the others. For Writing, Judge 10 was a bit more inconsistent than the others, as was Judge 3 for Vocabulary. In Structures, Judge 11 was much more inconsistent than the others. In short, it is clear that judges are not consistently inconsistent across skills, and thus there is no case for removing any judge completely, but rather on a skill-by-skill basis. Note that the table tells us nothing about agreement amongst raters – see below. However, on the evidence of the table alone, it would be sensible to remove Rater 12 from Listening, Rater 11 from Structures, and possibly Rater 4 from Reading, Rater 10 from Writing and Rater 3 from Vocabulary.

Such data are available for the three languages (Finnish, English, Spanish) rated in Version 1, SS. However, it is rather unclear how small inconsistencies should be compared with greater inconsistencies, and so, although the differences across judges above are evident, a more satisfactory, and conventional, way of judging intra-rater reliability is simply by correlating the results of a first rating with those of a second rating. Version 2a DESS did not allow any estimates of intra-rater reliability, and so this was only possible with Version 2b, STSE. Tables 6.3 and 6.4 give examples of the results, for German and French.

Table 6.3 Intra-rater reliability, German

Rater	Reading	Writing	Listening	Grammar	Vocabulary
R1	.92	.80	.91	.80	.86
R2	.88	.81	.84	.68	.88
R3	.92	.95	.87	.85	.91
R4	.83	.40	.86	.84	.90
R5	.69	.77	.90	.61	.86
R6	.92	.92	.99	.73	.94
R7	.95	.76	.91	.91	.92

Table 6.4 Intra-rater reliability, French

Rater	Reading	Writing	Listening	Grammar	Vocabulary
R1	.62	.67	.52	.72	.73
R2	.66	.68	.65	.87	.92
R3	.63	.22	.56	.78	.76
R4	.87	.86	.64	.89	.95
R5	.59	.71	.68	.97	.92
R6	.88	.82	.43	.95	.95
R7	.79	.76	.65	.83	.89

For Reading in German, the correlations are impressively high, with the exception of Rater 5, who is fairly inconsistent. In Writing, however, there is much more variability in the intra-rater reliability, with Rater 4 being unacceptably inconsistent at .4. In Listening, there is impressive agreement.The reliability for Grammar is on the whole lower than for Writing, but whereas the latter had one unacceptable rater, Grammar has three that are marginally, but not very, unreliable (at .6/.7). Finally, the intra-rater reliabilities for Vocabulary are entirely acceptable, with no dubious rater.

Intra-rater reliability for Reading is rather low for 4 of the seven French raters, similarly for Writing, although one rater (Rater 3) is remarkably inconsistent at .22. For Listening the situation is equally poor, with unacceptably low reliabilities. Very interestingly, however, the French judges are very consistent when rating Grammar, which, of course, is not described or scaled in the CEFR. Similarly for Vocabulary, there is high intra-rater consistency, with the possible exception of Raters 1 and 3 who are somewhat lower than is normally acceptable. This suggests either that the French judges are less familiar with the CEFR for language skills, or that they are more comfortable making judgements about linguistic abilities, perhaps because French language instruction focuses on linguistic features rather than on language use. This, it must be stressed, is mere speculation.

The only language where two different standard-setting procedures were used and can therefore be compared was English. Some measure of intra-rater consistency is possible by comparing those raters who used both Version 1 and Version 2 of the standard-setting procedures. Table 6.5 shows the results for Grammar. The correlations are not very high, but since different procedures were being used, albeit on the same items, this is perhaps rather encouraging, especially considering that the skill being compared is Grammar, for which there was less inter-rater agreement (and no CEFR scales) in any case and the fact that the interval between the two ratings was some 15 months.

Comparable results for Reading are available for one rater only, who achieved a correlation of .80 between SS Version 1 and DESS Version 2. This is very encouraging.

Table 6.5 Rater comparisons, using two different procedures, English Grammar

	R1, DESS Version 2	**R2, DESS Version 2**	**R3, DESS Version 2**
R1, SS Version 1	.69		
R2, SS Version 1		.66	
R3, SS Version 1			.56

Table 6.6 Contrast of mean ratings, by rater, German Reading

		Mean	N	Std Deviation	Significance
Rater 1	1st rating	3.42	55	1.572	
	2nd rating	3.76	55	1.677	.000
Rater 2	1st rating	3.35	60	1.655	
	2nd rating	3.23	60	1.671	NS
Rater 3	1st rating	3.25	60	1.385	
	2nd rating	3.35	60	1.376	NS
Rater 4	1st rating	3.23	60	1.500	
	2nd rating	3.97	60	1.615	.000
Rater 5	1st rating	3.22	60	1.563	
	2nd rating	3.30	60	1.442	NS
Rater 6	1st rating	3.55	60	1.682	
	2nd rating	3.63	60	1.507	NS
Rater 7	1st rating	3.30	57	1.647	
	2nd rating	3.42	57	1.658	NS

Apart from correlating the ratings from the first and second rounds, it is possible to contrast the mean levels given to the items on the first and the second rating. Examples are given in Table 6.6, for German.

Table 6.6 shows that two raters (R1 and R4) were significantly different in their ratings on the second rating. Both rated the items higher the second time around. Whilst there was a tendency for four of the other five raters to rate items more highly the second time, the differences were not significant.

Rater agreement

It is also possible to look at how much the raters agree with each other, skill by skill and language by language. Although it would be tedious to give full details here, I present an example in Table 6.7, to illustrate the results and the method. The agreement among raters of German Listening items, first rating, is very high.

Table 6.7 Inter-rater reliability, German Listening, first rating

	R1	R2	R3	R4	R5	R6	R7
R1	1	.95	.91	.88	.93	.93	.96
R2	.95	1	.89	.86	.88	.89	.91
R3	.91	.89	1	.86	.92	.89	.94
R4	.88	.86	.86	1	.83	.84	.84
R5	.93	.88	.92	.83	1	.92	.92
R6	.93	.89	.89	.84	.92	1	.89
R7	.96	.91	.94	.84	.92	.89	1

Table 6.8 Pearson correlations for item CEFR levels, by procedure (k = 50), English Reading

	SS Version 1 11 raters	**SS Version 1 9 raters**	**DESS Version 2 4 raters**
SS Version 1 11 raters		.92	.96
SS Version 1 9 raters	.92		.90
DESS Version 2 4 raters	.96	.90	

Kendall's W (a measure of concordance across all raters) for 7 raters is .918 (sig .000).

Standard-setting procedures compared

It is important to consider how this variation in judgements affects the CEFR levels assigned to items. Since different times of rating (first or second), different judges doing the rating, and different procedures used to elicit the ratings can all result in different ratings, the important question is: do these differences have an effect on whether an item is labelled A1 or A2 (etc.), and if so, by how much? In Table 6.8 we present the results for English of two different standard-setting procedures, with unreliable raters present and absent.

What is clear from this table is that there is a very high correlation between item CEFR levels, regardless of standard-setting method. Dropping unreliable raters has a variable impact, depending on the skill being investigated, but the results are still very comparable.

Table 6.9 Correlation between item logit values and judged CEFR levels, English Reading, by procedure

	Item logit value	**SS (Version 1) 11 raters**	**SS (Version 1) 9 raters**	**DESS (Version 2a) 4 raters**
Item logit value	1	.92	.91	.95
SS (Version 1) 11 raters	.92	1	.92	.96
SS (Version 1) 9 raters	.91	.92	1	.90
DESS (Version 2a) 4 raters	.95	.96	.90	1

Empirical item difficulties and judged item levels compared

Finally, we can examine the relationship between empirical item difficulties and the judged CEFR levels, across procedures. Table 6.9 shows that this relationship is very close, thereby giving us confidence in our procedures.

And finally

What does all this have to do with diagnosis? It is highly relevant in a number of ways. Firstly, as pointed out at the beginning of this chapter, a test score on its own does not carry much meaning. If it can be related to the performance of other learners, then there is a relative interpretation that can be given to the score. But that score will be more meaningful if the level of ability of those other learners is known. And that implies standards: knowing what level has been achieved and what that level means in terms of other possible levels. The CEFR is a set of standards for language development in a broad sense, and thus knowing that a learner has reached X level on the CEFR scale of development is more meaningful than merely knowing a score.

This is what I called diagnosis at the macro level – being able to relate a learner's performance to some known and accepted standard of ability or performance. To be able to do this, however, as this chapter aimed to show, means being able to set standards – being able to interpret a score against a set of standards, which means knowing where the cut-scores fall. This is only possible if a rigorous process of standard setting has been undertaken, with appropriate judges, and using appropriate procedures both of data collection and of data analysis.

Secondly, the experience of standard setting illustrated in this chapter has hopefully also demonstrated that standards are not objective: they are the result of a series of subjective, albeit expert, judgements made by individuals whose judgements one is prepared to trust. Without adequate familiarization with the targeted standards, without adequate feedback to judges on initial judgements about learners' performances or about the difficulty of items, then judges are unlikely to reach satisfactory levels of agreeement. And without agreement to some degree, standards cannot be set and therefore diagnosis is, at best, difficult.

Thirdly, for adequate detailed diagnosis of performance – what I have called micro-level diagnosis – one needs to know something about the property of the items on which the performance was demonstrated, otherwise it will be difficult, if not impossible, to judge the quality of the performance. Thus one needs to know how difficult a task has been set, and how well it has proved to discriminate among learners. Such data comes from piloting, as described in Chapter 5. But in order to interpret performance on a task, one also needs to know what that task represents in terms of a hierarchy of development. Is the task something that somebody who is at A2 can already do or is it something that

somebody who has already achieved B2 would find difficult? Thus one needs to know what level a task can be said to be 'at', or, better, what stage of development a learner might have to be at before one can be confident that they can perform a given task adequately. Again, without standard setting, it is impossible to answer such questions.

To conclude, in this chapter I have discussed the reasons why standard-setting procedures are needed, if we are to interpret test scores meaningfully and appropriately. I have described in some detail the two different procedures used by DIALANG in order to arrive at cut-scores to differentiate among the different levels of the CEFR. I have also illustrated the results of the analysis of such procedures and shown how it is possible to achieve reasonable reliability of judgements, and that empirical item difficulties can correspond well with judged CEFR levels of items. However, constant training and monitoring of judges is essential if believable levels are to be assigned to test scores.

Chapter 7: The Vocabulary Size Placement Test

Introduction

As was described in Chapter 3, the Vocabulary Size Placement Test (VSPT) is used in Version 1 of DIALANG, partly in order to give users some information about the size of their vocabulary, quite independently of the CEFR, but mainly to help the system decide whether to give users an easy, medium or difficult test. Once users have entered DIALANG, chosen the language of administration of the system, and then decided which language and skill they wish to be tested in, they are next presented with the option of taking a VSPT in the test language, with instructions in the language they selected for administration (the ALS – Administration Language Selected). Once they have completed the VSPT, they are given immediate feedback on their performance, in terms of six different bands of ability, ranging from *very low* to *indistinguishable from a native speaker*. If they have opted to take a test in Reading, Writing or Listening, they are then offered the opportunity of self-assessing their ability in the language and skill they are to be tested on. The self-assessment statements – SAs – are presented in the ALS, but there are as yet no SAs for Vocabulary or Grammar.

If users have responded to both the VSPT and the SAs, then the two results are combined to decide which level of test they will be given. If users have only responded to one of the two placement procedures, then that result alone is used to make the placement decision. If no placement procedure has been selected by the user, then the Medium test is administered.

Construct

The VSPT in Version 1 of DIALANG consists of a list of 75 words, and the user's task is to say whether each word is a real word in the test language or not. Fifty of these words are real words in the language, 25 are pseudo- or non-words. The user is not told how many pseudo-words there are, and all words are presented in a randomized order. The user

receives one point for each word correctly identified as real or pseudo, i.e. a response 'Yes' to a real word and a response 'No' to a pseudo-word. 'Yes' responses to a real word are known as 'hits' and 'Yes' responses to a pseudo-word are known as 'false alarms'. 'No' responses to real words are known as 'misses' and 'No' responses to pseudo-words are known as 'correct rejections'.

Scores can be calculated in a number of different ways, as we will see later in this chapter, but the reported VSPT score in Version 1 is derived according to an algorithm known as the Δm score. Essentially this involves penalizing users for guessing, by adjusting the score for real words by the number of false alarms (pseudo-words identified as real words). The original algorithm actually produces a score 'Z' which is then multiplied by 1,000 to produce the score which is reported in the VSPT feedback on a scale from 0 to 1,000.

There are VSPTs for all 14 DIALANG languages: they were constructed by Paul Meara and associates at the Centre for Applied Language Study, University of Wales in Swansea. Unlike other Yes/No vocabulary tests constructed by Meara (in particular the EFL Vocabulary Test, which consists of a number of tests on six different levels that are based on word frequencies in written texts), the VSPT does not use word frequency as its basis, since such frequency lists are not available in all the DIALANG languages. Since the DIALANG tests are intended to be as parallel as possible, it was important that the construction principle be the same for all languages. Moreover, unlike the EFL Vocabulary Test, the DIALANG tests were mainly to be used for placement purposes, and could therefore not be too long. In addition, test construction was constrained by available resources and so a suitably economic placement method had to be devised.

Meara decided to select words from one form class only, and he chose verbs rather than any other form class because there are relatively fewer verbs than other lexical form classes in any language. Nouns are a large form class, and adjectives are often derived from nouns. In some languages this selection procedure resulted in identifying most of the verbs in the language (Meara, personal communication, 2003) and thus the sample for that language could be considered highly typical of the language. This is not the case for languages like German, Dutch, English, German or Swedish, which can produce many phrasal verbs (in the case of English, phrasal verbs were not eligible for the VSPT because only single words were required for the items). Moreover, in English, nouns or adjectives can easily become verbs. Thus, the decision to use verbs only was supported by the Project. Indeed, it is hard to imagine a better procedure that would have resulted in a suitably objective placement procedure that took little time but still gave potentially meaningful results.

The real words were selected from medium-sized (not pocket) bilingual dictionaries and a semi-random set of 1,000 lexical verbs (not

auxiliaries or modal verbs) was chosen. Some procedures of dictionary sampling have been criticised by Nation (1993), and these criticisms were taken into account in the construction of the VSPT. One or two words were selected from every page (usually the first verb at the top of the left-hand page), because in some cases, for example German, Dutch and Swedish, there might be many verbs on one page (for example, verbs beginning with prepositional particles). This selection procedure nearly exhausts the available items in some languages and avoids the risk of weighting long entries, which sampling by page numbers could do.

A subset of 100 verbs was then selected at random from the 1,000-verb set.

Pseudo-words are defined by Huibregtse *et al.* (2002) as '*words that fulfil the phonological constraints of the language but do not bear meaning*' (p. 227). The false words were mostly generated specifically for this Project (although for some languages, some pseudo-words were taken from a set of pseudo-words often used in research in Swansea; see also Meara and Buxton, 1987). A linguistically aware native speaker was approached for each language and they were asked to produce a list of imaginary verbs. Pseudo-words bear no known relationship to frequency lists.

Once constructed, the draft 150-item tests (100 real words, 50 pseudo-words) were reviewed by native-speaker informants with instructions to check that the pseudo-words did not exist. As a result, some items were modified.

Piloting

The draft 150-item VSPTs were then incorporated into the DIALANG pilot system. A shot of one screen of the pilot English test is shown in Figure 7.1.

The purpose of the pilot was to examine the value of the VSPT and in particular to identify suitable items for inclusion in Version 1. Users were not given any choice: they could not choose to skip the VSPT, for obvious reasons, nor the SA statements, and they could not choose which skill to be tested on – they were simply presented with a test booklet which contained 50 items: 30 from one skill and 20 from either Grammar or Vocabulary. These were not presented in order of difficulty as difficulty had yet to be determined.

Comments from participants in and administrators of the DIALANG pilot tests suggested that the 150-item VSPT was too long as a placement instrument at the start of the test, even for piloting purposes. Discussions and trial analyses with different lengths of test led to the decision that a shortened, 99-item VSPT would be piloted for most languages, selecting 66 real words and 33 pseudo-words randomly from the original 150-word sample. After piloting, the VSPT contained in

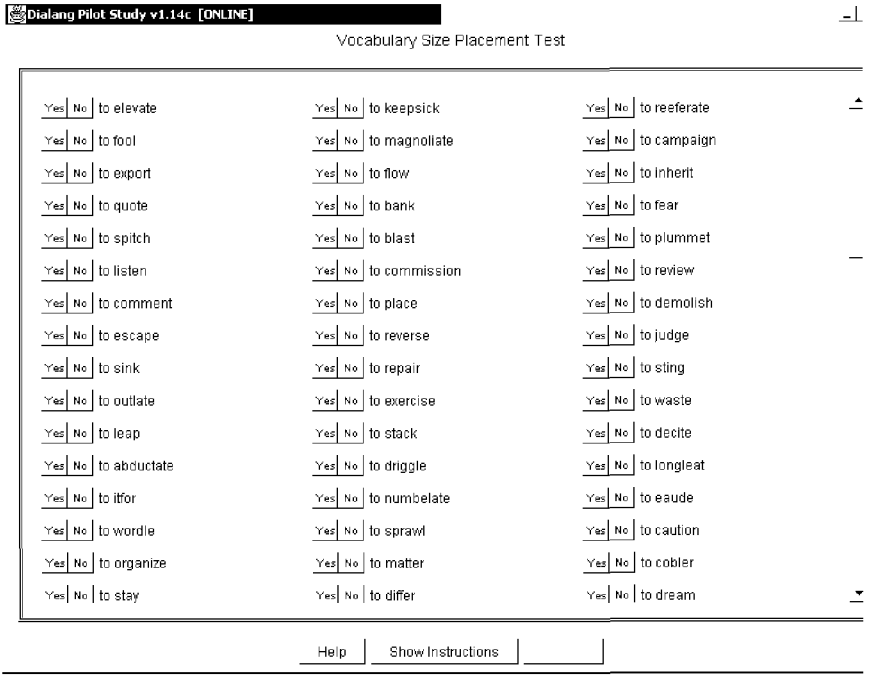

Figure 7.1 The pilot Vocabulary Size Placement Test

Version 1 of the system should consist of 75 words (50 real words, and 25 pseudo-words).

However, piloting of the English VSPT was completed before this decision was taken, and so the original 150 words were analysed, and a 75-item VSPT was constructed for Version 1 (50 real words and 25 pseudo-words). Normal item selection criteria were thought to be inappropriate, since the pilot test had been constructed using random selection procedures. In order to respect the sampling procedure, whilst at the same time shortening the test to manageable lengths, first a classical item analysis was made of all 150 items. Within the categories of real words and pseudo-words, the items were ranked in ascending order of difficulty, and then for each pair of items in the rank order, the one with better discrimination values was selected for inclusion in the shortened version, which became VSPT Version 1 (English).

Results for English

Descriptive statistics

The analysis of the responses of over 2,000 individuals to the 150 items of the English pilot VSPT showed that the test was relatively easy (mean

Table 7.1 Item analysis of VSPT pilot items

	Facility	Discrimination
Mean of pilot real words	82%	.346
Mean of pilot pseudo-words	79%	.225
Minimum (real)	33.0	.019
Minimum (pseudo)	45.7	.011
Maximum (real)	98.9	.514
Maximum (pseudo)	95.7	.330
SD (real)	15.048	
SD (pseudo)	12.597	
N	2094	
Cronbach alpha for pilot real words		.948
Cronbach alpha for pilot pseudo-words		.923
Overall pilot VSPT Alpha		.932

80.7 per cent), but had high reliability (.93). Table 7.1 shows the detail for real and pseudo-words separately.

The real words in the pilot test were easier than the pseudo-words (82 per cent vs 79 per cent) and discriminated somewhat better than pseudo-words. They were also slightly more reliable, but even the reliability for pseudo-words was very high. When scores on real words are compared with scores on pseudo-words, there is no significant correlation between the two ($r = -0.045$).

Version 1 of the VSPT, which consists of only 75 items, is still highly reliable (.9). If we compare the 75 items deleted from the VSPT with those that were selected for inclusion in Version 1, we find two closely parallel tests. The 75 items omitted from Version 1 correlated with Version 1 items at .878 which suggests a high degree of internal consistency in what the VSPT measures.

The comparative descriptive statistics are shown in Table 7.2.

Although the mean for Version 1 is only marginally higher than the omitted item mean, the scores are significantly different from each other ($p < .000$), since the standard deviation of Version 1 items is

Table 7.2 Comparison of Version 1 VSPT with omitted items

	N	Minimum	Maximum	Mean	Std Deviation
Items omitted from Version 1	2075	13.00	75.00	60.4786	7.67775
Items retained in Version 1	2075	13.00	75.00	60.8405	9.18376

notably higher, because of their greater discrimination. The reliability of the omitted items was .843, whereas for Version 1 items it was .895, again the result of somewhat higher discriminations in Version 1. The Project had thus been successful in creating a Version 1 VSPT of optimal quality.

Given the multilingual nature of the pilot populations, as we saw in Chapter 5, it is of interest to see whether the VSPT is easier for certain

Table 7.3 Real words correctly identified, by mother tongue

Mother Tongue	N	Mean %
Icelandic	71	90
German	450	87
Danish	115	86
Norwegian	103	86
Swedish	43	86
Dutch	252	83
Italian	27	81
Other L1	315	80
Finnish	349	79
French	160	77
Portuguese	40	77
Greek	11	72
Spanish	137	71
Irish	2	27

Table 7.4 Pseudo-words correctly identified, by mother tongue

Mother tongue	N	Mean %
Icelandic	71	86
Danish	115	84
French	160	84
Greek	11	84
Spanish	137	84
Norwegian	103	82
Portuguese	40	82
German	450	81
Italian	27	81
Swedish	43	81
Dutch	252	77
Finnish	349	75
Irish	2	72
Other L1	315	72

language groups. Turning back to the pilot VSPT of 150 items, Tables 7.3 and 7.4 present the results of real words and pseudo-words, for test-takers with different mother tongues.

Real words were easiest for speakers of languages related to English, notably Icelandic, German, Norwegian, Swedish and Danish, but curiously not for Dutch. In contrast, although pseudo-words were also easiest for speakers of Icelandic, speakers of German, Dutch and Swedish did comparatively poorly in comparison with speakers of French and Spanish. There is no obvious linguistic reason for this result, although speakers of languages other than the DIALANG languages (many of whom were speakers of non-Indo-European languages) did particularly poorly at identifying pseudo-words.

Scoring methods

The scoring procedure currently implemented in Version 1 of DIALANG corrects for false alarms (pseudo-words identified as real words). However, performance on real words seems to differ from performance on pseudo-words. The correlation between real words and total VSPT score is high (.84), whereas for pseudo-words it is only moderate (.51, which is partly a correlation of the pseudo-word score with itself). It could be that the ability to identify real words is somewhat different to the ability to identify pseudo-words. It may therefore be necessary to treat pseudo-word scores somewhat differently from scores on real words. It was thus decided to explore various possible scoring procedures.

A number of different scores for the VSPT were created. The first score, 'simple total', is a simple total of words correctly identified as either pseudo-words or real words (pseudo-words answered No are counted as correct).

Secondly, based upon a suggestion by Meara, a score called 'simple correction' was created by doubling the number of false alarms (pseudo-words wrongly identified as real words) before subtracting them from the sum of correctly identified real words ('hits').

Thirdly, a variable called 'raw hits' was created, which simply ignores pseudo-words and only gives credit for real words correctly identified.

In addition, three scores were produced by the Meara algorithm. One, 'Meara algorithm', is the raw results of the application of the Δm algorithm. A second, 'numerical Meara', rounds minus scores to zero and converts all other scores onto a scale of 0–1,000. This score is the numerical score reported to users in the VSPT feedback in Version 1. A third score – 'Meara Band' – is a conversion of 'numerical Meara' into six bands (where 1 is low and 6 is high), using the cut-offs for each band as reported to users. This is the band score reported to users.

The descriptive data for the various VSPT scores (including the Meara scores), based on the 150-item pilot test, are in Table 7.5.

Table 7.5 Descriptive statistics for pilot VSPT 150 item scores

	N	Minimum	Maximum	Mean	Std Deviation
Simple total	2059	29.00	150.00	121.29	16.34
Simple correction	2059	−64.00	100.00	60.78	22.12
Raw hits	2059	3.00	100.00	81.79	14.11
Meara algorithm	2059	−8.48	1.00	.49	.46
Numerical Meara	2059	.00	1000.00	527.52	278.76
Meara Band score	2059	1.00	6.00	3.9480	1.43

The highest mean of scores that take account of pseudo-words is gained by the Simple total score, and, of course, scores corrected for false alarms are lower than uncorrected scores ('Raw hits'). Correction can result in negative scores.

In order to see how these scores relate to each other, the intercorrelations of the various scoring methods are presented in Table 7.6. The lowest relationships are between Raw hits and all but the Simple total (which in any case includes Raw hits). The low correlation between Simple correction and Raw hits (.61) may be because the former score results in negative scores as well as positive ones. This supposition is reinforced by the relatively high correlation of .87 between the two scoring methods that deliver negative scores (Simple correction and Meara algorithm).

The closest relationships are between Simple total and Simple correction (.94), and Numerical Meara and Simple correction (.97). The close

Table 7.6 Correlations of pilot VSPT scores (n = 2059)

	Simple total	Simple correction	Raw hits	Meara algorithm	Numerical Meara	Meara Band
Simple total	1	.94	.84	.84	.91	.88
Simple correction	.94	1	.61	.87	.97	.94
Raw hits	.84	.61	1	.57	.59	.57
Meara algorithm	.84	.87	.57	1	.79	.78
Numerical Meara	.91	.97	.59	.79	1	.97
Meara Band	.88	.94	.57	.78	.97	1

relationship between Simple total and Simple correction is of course because both take account of performance on pseudo-words.

What is notable are the high correlations between the various Meara scores, using the complex Meara algorithm, and simpler procedures, including the Simple total and the Simple correction. It looks, *prima facie*, as if the extra complication of the use of the Meara algorithm might not be worth the effort. This needs to be confirmed, however, by examining the relationship between these scores and language proficiency scores – see next section.

Relationship between VSPT and test scores

Since the VSPT is used to predict the ability of users in order to present a test of suitable difficulty, we need to explore the relationship of the various scoring methods with criterion variables, namely the various language proficiency scores, arrived at through IRT calibrations. This is important not only in order to examine the VSPT's practical utility but also to understand its construct.

If we look at the relationship between the VSPT scores and the total scores for each of the five macro skills, we get the results set out in Table 7.7.

The best correlation between VSPT and the language tests is always with the Simple total scoring method, and the Simple correction score correlates almost identically as the Numerical Meara score (the score currently used in DIALANG). The lowest correlations are almost always with the Meara algorithm and the Raw hits score. The Simple total is clearly superior to a simple count of real words correctly identified (Raw hits). In other words, taking account of performance on pseudo-words is worthwhile. But there seems to be little justification for correcting scores by adjusting for false alarms (pseudo-words wrongly identified as real words) since Simple correction and the Numerical Meara scores always correlate lower than the Simple total with criterion variables. Scoring methods other than the Simple total score do not appear to add any value.

From a practical point of view, the VSPT, when scored by the Simple total method, appears to be a very useful predictor of performance on

Table 7.7 Correlation of VSPT with language tests

	N	Simple total	Simple correction	Raw hits	Meara algorithm	Numerical Meara	Meara Band
Reading	718	.64	.61	.53	.40	.63	.60
Grammar	1084	.64	.59	.54	.43	.60	.57
Writing	735	.70	.64	.59	.60	.65	.61
Listening	606	.61	.57	.52	.54	.57	.53
Vocabulary	975	.72	.66	.63	.62	.66	.62

the various language tests, and thus its function as a placement procedure for the main tests appears entirely justified. Vocabulary size as a predictor of other test scores – which is the point of the VSPT – seems to be best measured by a score which is a simple total of true words correctly identified, and pseudo-words correctly identified as false. There is no added value from any correction formula, but ignoring false words (the 'Raw hits' score) results in notably worse prediction. The extra effort involved in scoring with the Meara algorithm appears not to be justified, although even scores reported on a 6-band scale correlate quite respectably with criterion variables.

Despite the lack of correlation between pseudo- and real-word scores reported earlier, a score based on both real and pseudo-words proved superior to a score based solely on real words correct in predicting criterion variables. Thus, despite the lack of an intercorrelation between pseudo- and real-word scores, it would appear that taking account of pseudo-word performance adds considerable value to the VSPT.

The correlations of the VSPT with the various language tests showed the VSPT to be most closely related to the Vocabulary test, unsurprisingly, but only marginally less to the Writing test. However, since the Writing test is an indirect measure of writing ability and includes many lexically focused items, this is perhaps less surprising than appears at first sight. Although one might have expected the correlations between Vocabulary Size and Vocabulary Quality to have been higher, the correlations of Vocabulary with the Simple total are good, accounting for 50 per cent of the variance. The VSPT predicts Reading and Grammar with equal accuracy (.64) and predicts Listening somewhat worse (.61).

What this would appear to show is that the size of one's vocabulary is relevant to one's performance on any language test, in other words, that language ability is to quite a large extent a function of vocabulary size. From the point of view of the diagnosis of language strengths and weaknesses, a measure of vocabulary size would therefore appear to be of considerable value in its own right, let alone as a quick and reliable placement procedure for more detailed diagnoses of different aspects of language ability. This is in contrast to the opinion of Hughes (2003: 179) reported in Chapter 1, namely that tests of vocabulary were thought unlikely to yield useful diagnostic information.

In sum, there is good evidence for the usefulness of the VSPT as a predictor of performance on the various language tests, thereby justifying its use as a placement procedure, and at the same time providing some insight into the constructs of language ability.

Exploring the construct – subskills

In order to explore in more detail the construct of the VSPT, the relationship was examined between the Simple total score, which had performed best out of all the scores examined, and the various subskills of

the language tests. The following patterns were found and are detailed in Tables 7.8–7.11.

Table 7.8 Reading subskills

	Identifying main idea	Understanding specific detail	Inferencing
Simple total	.50	.47	.58

Table 7.9 Writing subskills

	Accuracy	Register/ appropriacy	Textual organization
Simple total	.70	.57	.51

Vocabulary size seems to relate most closely to the ability to infer from text, and somewhat less to the ability to understand specific details.

Interestingly, vocabulary size relates notably to Accuracy (of grammar, vocabulary and spelling) and considerably less to Textual organization (cohesion/coherence). An inspection of the items testing Accuracy may well show a preponderance of lexically focused items. One might have expected vocabulary size to relate more to Register/appropriacy than it does, since presumably at least part of the ability to produce or recognize appropriate language is a lexical issue, as is register.

Unlike the results for Reading, it appears that vocabulary size is slightly more related to the ability to identify the main idea when listening than it is to inferencing when listening. Certainly, the ability to understand detail bears relatively little relationship to vocabulary size, as with Reading.

Table 7.10 Listening subskills

	Understanding specific detail	Inferencing	Identifying main idea
Simple total	.44	.56	.60

Table 7.11 Vocabulary subskills

	Combination	Word formation	Meaning	Semantic relations
Simple total	.52	.60	.62	.56

'Combination' is described in the DIALANG Assessment Specifications as follows: '*This section focuses on assessing the knowledge and use of word meanings in terms of collocation and idiomaticity*'. 'Word formation' is described as '*This section focuses on assessing the knowledge and use of word meanings in terms of compounding and affixation*'. 'Meaning' is described as '*These vocabulary items focus on assessing the knowledge and use of word meanings in terms of denotation, semantic fields, connotation and appropriateness*'. 'Semantic relations' is described as '*This section focuses on assessing the knowledge and use of word meanings in terms of synonyms/antonyms/converses, hyponyms/hypernyms, and polysemous meanings*'.

The VSPT showed the closest relationship with Meaning in terms of denotation, semantic fields and connotation, and with Word formation, and it related somewhat less closely to collocation, idiomaticity and Semantic relations. Thus the VSPT could be said to be testing a combination of word formation and core meaning rather than semantic relations and extended meaning.

In order to explore this a little more, correlations were calculated of the different scoring procedures with the subskills of the Vocabulary test. This time a separate score was calculated for the number of pseudo-words correctly identified – see Table 7.12.

Interestingly, the Raw hits score results in better correlations than it achieved on all other criterion variables, which may suggest that a simple count of real words identified may have more to do with aspects of vocabulary knowledge than with other subskills. Nevertheless, a Simple total score, taking account of pseudo-words correctly identified, still performs best of all the scoring procedures on each subskill. However, the pseudo-words correct score shows markedly less relationship with any of the subskills, including Word formation, to which one might have expected it to relate rather more closely. It is still rather unclear what the pseudo-word score alone measures.

To throw some light on this, an exploratory factor analysis was conducted of the 75 VSPT items in Version 1. This resulted in 11 factors. Two factors appeared to be pseudo-word factors, and were separate

Table 7.12 Correlation of various VSPT scores with Vocabulary subskills

	Simple total	Simple correction	Raw hits	Meara algorithm	Numerical Meara	Meara Band	Pseudo-words correct
Combination	.53	.47	.49	.45	.47	.45	.21
Word formation	.60	.53	.56	.51	.54	.51	.24
Meaning	.62	.58	.53	.56	.59	.56	.33
Semantic relations	.56	.51	.49	.50	.51	.49	.26

from five other factors which were associated with real words. (The remaining factors related to small groups of words and defy explanation.) Although the results are not entirely clear-cut, it appears that the VSPT construct does indeed divide into real word and pseudo-word factors, and taking account of performance on pseudo-words enhances the ability of the VSPT to predict other language abilities.

Analysis of the relationship between the VSPT and background variables

The reader will remember from Chapter 5 that data were gathered during the piloting on test-taker characteristics. It is therefore possible to examine the relationship between performance on the VSPT and these background variables. This was done using the Simple total VSPT score. The results for real words and pseudo-words were given above. Below I give the results for the combination, in Version 1 (75 items).

Analysis of variance showed highly significant differences among means for the various mother tongues ($F = 22.014$, $p < .000$), shown in Table 7.13. *Post-hoc* contrasts showed that Icelandic speakers perform best on the English VSPT, followed by Danish, Norwegian, German and Swedish speakers. This could possibly be due to the six languages being related. Spanish speakers perform worse even than those whose mother tongue is not a DIALANG language. However, this could be due to higher language proficiency in those groups scoring highest.

After adjusting for language proficiency through an analysis of covariance, there was indeed a significant effect of mother tongue on vocabulary size. However, this accounted for only 4.8 per cent of the variance on the VSPT. There was a strong relationship between

Table 7.13 Mean scores on VSPT (by mother tongue)

Mother tongue	N	Mean %
Icelandic	71	89
Danish	116	86
Norwegian	103	86
German	450	85
Swedish	43	85
Dutch	252	82
Portuguese	40	80
Finnish	349	77
Italian	27	77
Other L1	315	77
French	160	75
Greek	11	75
Spanish	137	75
Irish	2	41

Table 7.14 VSPT mean scores adjusted for language proficiency

Mother tongue	Mean %
Icelandic	85
Danish	84
Norwegian	83
Swedish	83
Portuguese	82
Italian	81
Dutch	80
French	80
Other L1	80
Spanish	77

* For technical reasons, the German, Greek and Irish groups had to be removed from this analysis

language proficiency (the covariate) and the VSPT scores (31 per cent of the variance in the VSPT score is explained by language proficiency). Thus, adjusting scores for language proficiency, Icelandic speakers have the highest vocabulary size score, followed by Danish, Norwegian and Swedish. This must be due to the relatedness of these languages. But language proficiency is a much greater determinant of VSPT scores than a test-taker's mother tongue.

Table 7.14 gives the adjusted mean scores for each group.

This clearly indicates the influence of mother tongue on a measure of vocabulary size, even though this influence is much less than that of language proficiency.

With respect to level of education, university students score highest, and those with only a primary level of education score worst (Table 7.15).

However, there are no significant differences between non-university higher, and secondary general levels of education, although secondary vocational is significantly lower than both (and no higher than primary level).

Table 7.15 VSPT scores by education

Education	N	Mean %
Higher (university)	1040	84
Higher (non-university)	287	80
Secondary (general)	423	80
Secondary (vocational)	175	73
Primary	89	73
Other	45	83

A one-way analysis of variance showed highly significant differences among means for length of time learning English. Those who have been studying English for more than 10 years score the highest on the VSPT. The difference between those learning English for less than a year and those learning the language for between 1 and 2 years is not significant. Otherwise, the VSPT mean score increases as expected by number of years English has been studied, until 7–8 years, where there is then no difference in VSPT mean scores between this group and those who have been studying English for 9–10 years or 5–6 years. See Table 7.16.

Table 7.16 VSPT scores, by length of time learning English

How long	N	Mean %
Less than a year	62	74
1–2 years	113	69
3–4 years	218	76
5–6 years	427	80
7–8 years	415	82
9–10 years	410	82
More than 10 years	431	88

Regarding frequency of use of English, the highest VSPT scores are achieved by those who use English nearly every day (Table 7.17). Otherwise there are no significant differences between the groups, which is somewhat surprising. It would appear that vocabulary size is independent of frequency of language use, and more related to how long one has studied the language.

There is no significant difference between mean scores on the VSPT between males and females (Table 7.18).

Table 7.17 VSPT scores by frequency of use of English

How often	N	Mean %
(Almost) every day	526	87
Once or twice a week	913	80
Once every two weeks	177	79
Once a month or less often	267	77
Cannot say	193	78

Table 7.18 VSPT score by sex

Sex	N	Mean %
Female	1271	81
Male	805	81

Although there are significant differences in VSPT scores by age overall ($F = 5.412$, $p < .000$), the only significant contrasts are between the Under 18s and the 26–35, 36–45, 46–55 and 56–65 age groups, with the 56–65 age group scoring highest and the Under 18 age group scoring lowest. Results are presented in Table 7.19.

Table 7.19 VSPT score by age

Age	N	Mean %
Under 18	58	75
18–25	1295	81
26–35	372	82
36–45	208	83
46–55	119	84
56–65	20	88
Over 65	4	84

Table 7.20 Self-assessed CEF level

Overall SA level	N	Mean
A1	173	71
A2	213	70
B1	538	78
B2	595	83
C1	325	88
C2	232	92

In addition to responding to a background questionnaire, test-takers also self-assessed their language ability in the skill being tested. Table 7.20 shows how their overall self-assessment in terms of the Council of Europe Framework levels related to their VSPT scores.

Reassuringly, there are highly significant differences in VSPT scores, by self-assessed CEFR level ($F = 175.153$, $p < .000$). However, there is no significant difference between those rating themselves A1 and A2 in their VSPT score.

Yet the correlation between self-assessed CEFR level and a respondent's VSPT level is a moderate .483 (Spearman rho). This is notably lower than the correlations, shown in Table 7.7 above, of the VSPT with the test scores (ranging from .61 to .72).

Conclusion and further research needed

The relationship between vocabulary size and language proficiency, by skill, has been shown to be substantial for English. The question of the value of a test of vocabulary size has thus been answered positively.

However, the relationship between vocabulary size and self-assessed levels on the Common European Framework has been shown to be quite modest. It is therefore clear that the two placement procedures used in DIALANG to 'pre-estimate' a candidate's probable level of ability are complementary. Certainly the VSPT predicts language proficiency rather well.

Clearly, similar studies need to be conducted for the other 13 languages in the DIALANG system, once sufficient data become available.

In terms of diagnosis, it would appear that a test of vocabulary size, even one that is not based upon word frequency studies, relates substantially to more traditional measures of language proficiency. Although the VSPT is not based upon the CEFR, this does not *a priori* preclude it from being a potentially useful diagnostic measure, and the results presented in this chapter suggest that there could indeed be diagnostic value in a discrete measure of lexical competence, of a fairly basic nature. This adds weight to the hypothesis presented in Chapter 1 that useful diagnostic tests may well be more discrete-point than integrative. It may be useful to focus upon basic aspects of linguistic competence, like vocabulary size, when seeking to diagnose a learner's foreign language proficiency, however unfashionable such a test might be, and despite the apparent lack of relationship between it and the Council of Europe's Common European Framework. If such a test appears to be useful, as indeed it does, then further research is necessary to understand better what constructs underlie it.

It would be valuable to compare performance on this Yes/No test with performance on other vocabulary measures, both those based on word frequency counts, like the Vocabulary Levels test, and tests of productive vocabulary knowledge (see Laufer and Nation, 1999, and Read, 2000).

It has been suggested (Paul Meara, personal communication) that an external criterion for validating the VSPT could be a parallel translation task: once users have responded Yes or No to an item, they could be asked to give a translation equivalent in their L1.

It would be worth exploring the performance of the VSPT at different levels of language proficiency to see whether users at higher and lower levels of language proficiency perform differently on the test. Is there, for example, a difference between high and low proficiency users in their performance on pseudo-words? Does the frequency of false alarms across high and low proficiency learners differ?

Given the lack of correlation between performance on real words and pseudo-words, it would be worth investigating how users respond to these two different item types. One possibility would be to conduct a concurrent think-aloud study into the process of rejecting and accepting pseudo-words, and another possibility would be to conduct a series of interviews with students to see what they say about what is being tested on both real and pseudo-words

This research has shown that the parallel forms' reliability of the test is very high. However, we have yet to investigate the test–retest reliability, which would be worth doing.

Finally, all the above suggestions should be replicated for all 14 DIALANG languages, since the system is a unique and valuable resource for the investigation of aspects of language proficiency.

Chapter 8: The role of self-assessment in DIALANG

Self-assessment is a central component of DIALANG, because it is believed to be central to language learning (Holec, 1979; Oscarson, 1989; Little, in press). The DIALANG philosophy of diagnosis is that a comparison is essential of one's self-assessment in any given aspect of language use with one's performance on a diagnostic test which is not high-stakes, and which therefore risks minimal distortion of one's ability. Such a comparison may reveal interesting discrepancies between one's self-assessment and the test result, and may therefore be useful for insights into one's learning and one's beliefs about one's abilities.

As described in Chapter 3, self-assessment is integral to DIALANG in various ways. First, users are offered the opportunity to self-assess their abilities in Reading, Listening or Writing, before they take a test in one of these skills. The information from the self-assessment is used, either in combination with the VSPT if they have also taken the VSPT, or alone, to assign users to tests at one of three levels of difficulty. The users' self-assessment is also compared with their eventual test score, in order to give users feedback on the correspondence between the two. In addition, users are offered the opportunity to browse a series of possible explanations for any mismatch between self-assessment and the test results, and so self-assessment is also an integral part of the feedback system, as well as of the tests proper.

Chapter 5 has briefly described the piloting process and given some indication of the self-assessed levels of proficiency of the pilot sample. In this chapter, we will look in more detail at the self-assessment component and what the pilot results show about its value. After a description of the self-assessment system in more detail we will look at the 'I Can' statements themselves and what the empirical results reveal, and then we will look at users' self-assessment results, their relationship to language proficiency and to background variables.

Description of self-assessment statements

DIALANG is closely based on the Common European Framework, and in particular on the global and illustrative scales that are at the core of the operationalization of the CEFR. These scales are couched in terms of descriptions of what learners can do in various aspects of language use at the different levels of the CEFR. Appendix B to the CEFR gives a detailed description of how these scales were devised and calibrated. In essence, scales were assembled from a wide variety of international sources, and deconstructed into descriptive content categories to result in a pool of edited descriptors. These descriptors were then presented to teachers to sort into the categories they thought they described. The teachers then judged the adequacy and clarity of the descriptors and sorted them into bands of proficiency. These proficiency scales were then subject to a number of quantitative analyses, as described in Appendix B of the CEFR and in North (2000), in order to arrive at sets of scaled descriptors in various areas of language proficiency.

CEFR Appendix B (pp. 222 and 223) lists these areas, which range from understanding interaction between native speakers to listening to radio and audio recordings, to reading correspondence and instructions. In the productive skills the areas range from taking part in formal discussions in meetings to interviewing and being interviewed, delivering sustained monologues and addressing audiences and, in written mode, from creative writing to writing reports and essays.

Given the constraints on computer-based testing, the CEFR scales for spoken production were ignored, and DIALANG concentrated on developing scales for Reading, Listening and Writing. Appendix C of the CEFR and Huhta *et al.* (2002) describe how the DIALANG self-assessment scales were developed. A Project team reviewed all relevant CEFR statements and selected those which were the most concrete, clear and simple. More than 100 statements were chosen in this way for further development. The wording was changed from 'Can do' to 'I can', since they were to be used for self-assessment, not teacher-assessment, and some statements were simplified for their intended audience. A small number of new statements was also created. After extensive review, the original statements were carefully translated into the 13 other languages of DIALANG, and then reviewed, revised and re-reviewed until the Project was satisfied that the statements were equivalent in all 14 languages.

This process resulted in 35 detailed statements for self-assessment of Reading, 33 statements for Writing and 43 statements for Listening. There were also 6 overall self-assessment statements for each of the three skills, which were piloted but not included in Version 1 of DIALANG. The results of the overall scales will also be described in this chapter. The reason they are not used in Version 1 – although they may be restored in later versions of DIALANG – is that it was believed

that self-assessment is easier and more accurate if the learner can refer to a concrete situation or task rather than using a more general scale. And, of course, the overall self-assessment is in effect a one-item instrument, whereas the detailed self-assessments contain many more items and so are more like a real test, and likely to be more reliable.

In addition, the Project also developed a set of concise descriptors for Reading, Listening and Writing, based on the CEFR, to accompany the test score. When learners receive feedback on their performance, they are given not only their result at a CEFR level, but also a brief description of what that level means. These 'reporting scales' were repeatedly validated by DIALANG in the standard-setting procedures (described in Chapter 6), where they were deconstructed into meaning units, and then reconstructed by expert judges into the six CEFR levels as part of the familiarization process of the standard setting.

It is worth pointing out here that both the CEFR original descriptors, and the Can-Do SA statements are generally positive, and only rarely is something negative suggested in a statement. Thus DIALANG does not contain any 'I cannot (yet) do . . .' statements, which might have added valuable information. Bachman and Palmer (1989) reported that their informants were somewhat better at evaluating what they had difficulty doing rather than what they could do, and so there might be a case for exploring the value of negatively worded statements for diagnostic purposes. We return to this point in Chapter 17.

Both the overall scales and the detailed self-assessment statements were pre-tested and subsequently analysed and calibrated during the piloting process, and this chapter examines some of those results. Before doing so, however, it is worth mentioning a further development of Can-Do statements undertaken by DIALANG which will be relevant to diagnosis of strengths and weaknesses in due course, and these are the linguistic Can-Do scales.

As has been mentioned several times, the CEFR does not contain any information about linguistic proficiency in terms of specific languages – the CEFR is intended to be neutral as to language. Thus there are no self-assessment statements that relate to specific languages in the CEFR. However, DIALANG includes tests of vocabulary and of grammar in all 14 languages, and these tests needed to be related as far as possible to the CEFR, in order for the reporting of results to be consistent with the reporting of performance on the language use tests.

Therefore, as part of the development of standard-setting procedures as reported in Chapter 6, DIALANG developed sets of Can-Do statements for Vocabulary and Grammar, in overall scales at six levels parallel to the overall scales for the skills, and also broken up into meaning units similar to those used in the process of validating the overall scales for language use skills (35 units for English Vocabulary, 23 for English Grammar). The Project examined the CEFR closely, in particular those scales that relate to the assessment of linguistic ability on productive

tasks, especially but not only those contained in Chapter 5 of the CEFR on 'The users'/learners' competences'. These were then subject to an extended process of peer review and revision within the Project and the English versions were also sent for external review by applied linguists experts in second language acquisition and vocabulary development, as well as in the development of measures of vocabulary acquisition. The Assessment Development Teams for each language developed scales for their own languages in a parallel process, basing themselves on the CEFR, referring to the English scales as appropriate, and to any other sources they deemed relevant. Inevitably, these scales vary considerably across languages. For example, for French Vocabulary, this process resulted in 13 separate units, and for Grammar in 17 units, whereas for German there were 27 units for Vocabulary and 40 for Grammar.

These scales were then used in the standard-setting procedures for Vocabulary and Grammar for the different languages. Although they doubtless need further refinement and validation, it is significant that expert judges often reached greater agreement about the levels of units in these scales than they did with the language use skills (experts in French and Portuguese in particular showed notably greater inter-rater agreement for these scales than for the skill scales). These linguistic scales are thus potentially a source of further insights into users' self-assessed proficiency and might one day be of value for the diagnosis of abilities in vocabulary and grammar. However, the account of results in this chapter will concentrate on the self-assessment of Reading, Writing and Listening abilities.

Analysis of self-assessment statements

The piloting of the self-assesment statements took place in a complicated linked design of statements, which changed during the piloting. Consequently, the numbers of users reponding to any one statement varies considerably. If we simply take those self-assessment statements that were responded to by all pilot test-takers in any one booklet, the reliabilities vary somewhat (see Table 8.1), but are quite respectable, given the relatively small number of statements included, and the fact that many of the statements, especially at A1 and A2, showed very little variance (since most users claimed they could do these things).

Version 1 of DIALANG contains 18 self-assessment statements per skill (and these are the same regardless of test language). This number was decided on the basis of an inspection of content, the empirical 'difficulties' of the pilot statements, coverage across the six levels of the CEFR, and an estimate of how many statements users could bear to answer before fatigue or frustration set in. Table 8.2 shows the Cronbach alpha reliabilities of those piloted self-assessment statements that appear in Version 1.

Table 8.1 Cronbach alpha reliabilities of English self-assessment statements by skills and booklets

Skill/Booklet/subset	N cases	K of SA statements	Cronbach alpha
Reading	467	10	.713
Booklet 25	183	9	.743
Booklet 26	174	13	.785
Booklet 27	180	12	.732
Writing	472	11	.759
Booklet 29	182	11	.785
Booklet 30	186	12	.808
Booklet 31	185	13	.857
Listening	385	13	.786
Booklet 33	164	13	.756
Booklet 34	168	11	.752
Booklet 35	136	12	.751

Table 8.2 Reliability of self-assessment statements in Version 1, English

Skill	N of cases	K of SA statements	Cronbach alpha	Spearman-Brown prediction (if k = 45)
Reading	376	18	.813	.916
Writing	380	18	.874	.945
Listening	309	18	.823	.921

These reliabilities are entirely acceptable, especially in the light of the relatively small number of items in each set of statements. If we use the Spearman-Brown prophecy formula to predict what the reliabilities would be if more statements were added – say 45 statements, in order to be comparable both with test reliabilities and with the data in the next paragraph – then the predicted reliabilities would be .916, .945 and .921 respectively.

Phase 1 of DIALANG analysed 45 self-assessment statements (written in either Finnish, Swedish or English) across the three skills Reading, Writing and Listening, as delivered to takers of Finnish tests, and reported the correlation of their calibrated values with the original CEFR levels. This was .880 (Spearman) total and for the three skills:

Reading	.918
Writing	.915
Listening	.913

The correlation for the self-assessment statements delivered with the English tests in Phase 2 with the original CEFR levels was .930 (Spearman) total and for the three skills:

Reading	.928
Writing	.920
Listening	.911

Thus there are good reasons for believing that the DIALANG self-assessment statements correspond very closely to the original CEFR levels.

To examine to what extent the SA statements were equivalent across different languages of administrations (ALS – remember that users could choose in which language they wished to read the self-assessment statements), the difficulties of the 18 Version 1 statements were correlated for each ALS in the English data. The results for Reading are shown in Table 8.3.

Table 8.3 Intercorrelations of 18 SA statements, by ALS. Reading ability

	Danish	**Dutch**	**English**	**Finnish**	**French**	**German**	**Norwegian**	**Spanish**
Danish	1.000	.959	.961	.926	.946	.944	.910	.946
Dutch	.959	1.000	.960	.959	.962	.970	.912	.971
English	.961	.960	1.000	.927	.966	.951	.868	.967
Finnish	.926	.959	.927	1.000	.942	.960	.882	.948
French	.946	.962	.966	.942	1.000	.969	.879	.982
German	.944	.970	.951	.960	.969	1.000	.868	.988
Norwegian	.910	.912	.868	.882	.879	.868	1.000	.882
Spanish	.946	.971	.967	.948	.982	.988	.882	1.000

Table 8.3 suggests very strongly that the levels of the self-assessment statements are equivalent across languages. In addition, the Finnish logit values for each self-assessment statement were correlated with the English logit values and the results were .916 (Spearman) total, and for the three skills separately:

Reading	.901
Writing	.904
Listening	.979

Despite these impressive correlations, some self-assessment statements performed oddly in relation to the original CEFR levels, inasmuch as some had logit values higher or lower than the minimum or maximum logit values of the CEFR level they supposedly belonged to. This suggests that such statements should be at a higher or lower level than in the CEFR. It should, however, be remembered that the original scaling of the CEFR Can-Do statements was based on teachers' perceptions, whereas the data presented here is based on learners' own self-assessments.

In the English data, the number of statements that 'changed levels' were:

Reading	4/18
Writing	3/18
Listening	4/18

However, even if we accept the proposed new levels based on the English data, the correlation between the new self-assessment levels and the original CEFR levels is still .936.

The statements that it is suggested might need to have their level changed are as follows:

Reading

	New level	CEFR	Self-assessment statement
R23	B1	B2	I can read correspondence relating to my fields of interest and easily understand the essential meaning.
R27	C1	B2	I can go quickly through long and complex texts, locating relevant details.
R31	B2	C1	I can understand in detail a wide range of long, complex texts provided I can reread difficult sections.
R33	B2	C1	I can understand any correspondence with the occasional use of a dictionary.

Writing

	New level	CEFR	Self-assessment statement
W17	B2	B1	I can describe basic details of unpredictable occurrences, e.g. an accident.
W23	B1	B2	I can develop an argument giving reasons in support of or against a particular point of view.
W33	B2	C2	I can provide an appropriate and effective logical structure, which helps the reader to find significant points.

Listening

	New level	CEFR	Self-assessment statement
L04	B1	A2	I can understand enough to manage simple, routine exchanges without too much effort.
L15	A2	B1	I can follow clear speech in everyday conversation, though in a real-life situation I will sometimes have to ask for repetition of particular words and phrases.
L22	A1	B1	I can catch the main points in broadcasts on familiar topics and topics of personal interest when the language is relatively slow and clear.
L35	B2	C1	I can follow most lectures, discussions and debates with relative ease.

One question is to what extent are we justified in changing the levels of the self-assessment statements from the original CEFR levels? Although this is an interesting question, it probably does not make any difference to DIALANG, since users are assigned a self-assessed level based upon their responses to all 18 self-assessment statements, and since the correlations with the original CEFR levels are so high anyway. Nevertheless, it would be worth investigating to what extent the levels of self-assessment statements remain stable across different populations of users. Given that there were relatively few users in the DIALANG piloting of English at low levels of proficiency, and since most proposed changes of level involve a reduction in the level of the statement, it could be that these results will not be replicated with other populations, in particular with learners of less widely taught languages, or where the test-taking population has lower levels of proficiency.

Analysis of users' self-assessed scores

There are a number of ways of calculating a user's self-assessed level from their responses to the 18 detailed self-assessment statements. One is simply to give them a raw score based upon the number of self-assessment statements they respond to positively. A second procedure is to weight their responses to each statement by the CEFR level of that statement (where A1 is 1, A2 is 2, B1 is 3, and so on). In fact, the correlation between these two scores, raw and weighted, is very high:

Listening	.973 (n = 538)
Reading	.977 (n = 632)
Writing	.981 (n = 650)

It is also possible to create scores based on the IRT analyses of the statements, i.e. taking into account the logit value of the statements. First, an IRT-weighted raw score is calculated, which weights the response to any self-assessment statement by the discrimination value of that statement that is produced by the IRT computer program OPLM. This is then used in a look-up table to assign test-takers a self-assessment IRT 'score', and an overall CEFR level based upon cut-offs derived from a standard setting procedure involving four judges. Currently, DIALANG uses this latter method to assign self-assessed levels.

The intercorrelations of the self-assessment IRT 'score' with the raw and weighted raw scores are shown in Table 8.4.

Clearly the weighted SA score is superior to the raw SA score in its relation to the calibrated IRT-based 'score'.

Table 8.4 Correlation of self-assessment IRT 'score' with raw and weighted raw scores

	Raw SA score	Weighted raw SA score	N size
Reading	.763	.850	624
Writing	.810	.866	645
Listening	.795	.855	534

Intercorrelation of Overall with Detailed self-assessment statements

As it is in principle interesting to know whether a student's Overall self-assessment agrees with the result of the Detailed assessments – and indeed this may be a way of trying to validate a score derived from the detailed statements – it is worth trying to compare the two sets of self-assessments (SA).

The Spearman rank-order intercorrelations of Overall SA level and Detailed SA, for both raw and weighted scores, by skill, are as in Table 8.5.

Learners seem better able to self-assess (in the sense that their Overall self-assessment agrees better with their Detailed self-assessment) their

Table 8.5 Correlations of Overall SA with Detailed SA based on 18 Version 1 statements

Overall Skill	N	Detailed raw score	Detailed weighted score
Listening	538	.504	.561
Reading	632	.472	.528
Writing	650	.577	.617

Table 8.6 Correlation of Overall SA with Detailed IRT-based SA CEFR level

	CEFR level
SA Listening ability	.667 N = 534
SA Reading ability	.628 N = 624
SA Writing ability	.677 N = 645

productive skills than their receptive skills, presumably because they get feedback on the former and not the latter. In every case, the weighted self-assessment score correlates better with the Overall self-assessment than does the raw self-assessment score. This is evidence for the somewhat greater validity of the weighted score. In addition, a weighted score makes more sense conceptually than a simple raw score (since it takes account of the difficulty of the skill being assessed). However, a CEFR level arrived at through an IRT-based self-assessment parameter correlates even better with the Overall self-assessment (Table 8.6).

Although these results show a degree of correspondence between a single estimate of one's ability in a skill – the Overall self-assessment – and the result of a more detailed self-assessment arrived at by responding to a number of self-assessment statements, in point of fact Version 1 of DIALANG does not ask users to respond to an Overall self-assessment statement. Instead it prefers to ask users to self-assess in detail and the Overall self-assessment statement is used only in reporting the results of test performance. However, it may be that at some point in the future, an obligatory Overall self-assessment statement could replace the 18 Detailed self-assessment statements, which are currently optional in the DIALANG system. These optional statements could then be available in a separate self-assessment module for users to explore (and to receive detailed feedback on their self-assessment).

Self-assessed levels and test results

In this analysis, two variables are used to indicate self-assessed level: the Overall self-assessment, and the self-assessed IRT 'score' that resulted from the calibrations of the Detailed self-assessment statements. To indicate language ability by skill, the IRT-based logit score is used that resulted from the calibrations of results based upon an n size of 2060 (for English).

The Spearman correlations between Overall SA and test scores are as in Table 8.7:

Table 8.7 Correlation of Overall SA with calibrated skills test scores

	Reading	Writing	Listening	Grammar	Vocabulary
Overall SA	.544	.580	.474	.522	.554
N	718	735	606	1084	975

Table 8.8 Correlations between Detailed self-assessment IRT 'scores' and calibrated skills test scores

	Reading	Writing	Listening	Grammar	Vocabulary
Detailed SA parameter by skill	.487	.550	.495	Reading SA .504 Writing SA .439	Listening SA .548 Writing SA .595

What Table 8.7 shows is a somewhat modest correlation only between Overall self-assessment and performance on a test of that skill. 'Accuracy' – in the sense of the closest correlation – is achieved by self-assessment of Writing, and it may be reasonable to assume that this is because language users get more feedback on their productive abilities (in this case written performances) than they do on their Reading or Listening abilities. Interestingly, the correlations of Grammar and Vocabulary with self-assessment of one's ability in an unrelated area (.52 and .55 – remember that there are no self-assessment scales for Grammar or Vocabulary) are almost as good as the correlations of self-assessment of ability with a test of that ability (.54, .58 and .47).

With the exception of Listening, a Detailed self-assessment correlates lower with a test of the ability being assessed than does an Overall self-assessment (Table 8.8). However, a Detailed self-assessment of Writing or Listening abilities seems to correlate quite high with one's score on a test of Vocabulary (see Table 8.8). Quite what this might mean requires further research. However, it is arguably more important to look at the relations between CEFR levels as estimated by tests and self-assessments, rather than simply comparing 'scores', which we do in the next section.

CEFR levels according to self-assessment and according to test results

The above results have compared test scores and self-assessment 'scores'. However in both cases, 'scores' are converted into CEFR levels before being reported to users. It is thus of interest to know to what extent the CEFR level identified by a test result corresponds to a CEFR level from the self-assessment, and how many 'discrepancies' are identified. Tables 8.9–811 report such analyses for Reading, Writing and Listening.

Table 8.9 Self-assessment vs test score, Reading, by CEFR levels

		Current CEFR cut-off Reading						Total
		A1	A2	B1	B2	C1	C2	
CEFR Reading SA Detailed	A1	1	10	0	0	0	0	11
	A2	0	59	0	0	0	0	59
	B1	0	34	61	0	0	0	95
	B2	0	0	50	192	63	0	305
	C1	0	0	0	0	50	71	121
	C2	0	0	0	0	0	33	33
Total		1	103	111	192	113	104	624

Chi square p = < .000, Pearson r = .894, Spearman rho = .912

Table 8.10 Self-assessment vs test score, Writing, by CEFR levels

		Current CEFR cut-off Writing						Total
		A1	A2	B1	B2	C1	C2	
CEFR Writing SA Detailed	Below A1	1	0	0	0	0	0	1
	A1	1	16	0	0	0	0	17
	A2	0	27	42	0	0	0	69
	B1	0	0	120	207	76	0	403
	B2	0	0	0	0	32	0	32
	C1	0	0	0	0	8	81	89
	C2	0	0	0	0	0	34	34
Total		2	43	162	207	116	115	645

Chi square p = < .000, Pearson r = .857, Spearman rho = .839

Table 8.11 Self-assessment vs test score, Listening, by CEFR levels

		Current CEFR cut-off Listening						Total
		A1	A2	B1	B2	C1	C2	
CEFR Listening SA Detailed	A1	5	8	0	0	0	0	13
	A2	0	9	49	0	0	0	58
	B1	0	0	120	65	0	0	185
	B2	0	0	0	177	0	0	177
	C1	0	0	0	30	37	3	70
	C2	0	0	0	0	0	31	31
Total		5	17	169	272	37	34	534

Chi square p = < .000, Pearson r = .890, Spearman rho = .866

As can be seen, there are highly significant associations between SA and test CEFR levels. For Reading, the correspondence is close for A2, B1 and B2, whereas for C1 it is somewhat less. For Writing, there is less discrimination in self-assessment at B1, and at C1 in the test results, meaning that there are necessarily greater discrepancies between test results and self-assessment. For Listening, there is relatively little discrimination in test results, which cluster around B1 and B2.

Nevertheless, there is considerable agreement between test and self-assessment results in terms of CEFR levels, for all skills. Of course, discrepancies may be due to under- or over-estimation on the part of the learners, or to inappropriate cut-offs for the CEFR levels in either tests or self-assessments or both. Only further data collection and analysis will be able to resolve such uncertainties.

Self-assessment and background variables

It will be recalled from Chapter 5 that during the piloting of DIALANG a background questionnaire was administered to test-takers, as well as the self-assessment statements. This enables a study of the relationship between responses to the questionnaire and respondents' self-assessments. Here we compare the Overall self-assessment with various background variables – sex, age, education, length of time learning English and frequency of use of English. Mother tongue will be examined in the next section.

Sex

Table 8.12 compares males' and females' self-assessments.

Males rate themselves somewhat lower than females, although there is no statistically significant difference.

Table 8.12 Overall self-assessment by sex

Ses	N	Mean	Std Deviation
Female	1271	3.73	1.387
Male	805	3.57	1.372

Age

Table 8.13 shows that the youngest rate themselves lowest in terms of the CEFR and the oldest rate themselves highest.

Education

University students rate themselves highest, and primary students rate themselves lowest (see Table 8.14).

Table 8.13 Self-assessed mean scores by age

Age	N	Mean	Std Deviation
Under 18	58	3.14	1.290
18–25	1295	3.61	1.328
26–35	372	3.77	1.473
36–45	208	3.86	1.406
46–55	119	3.83	1.542
56–65	20	3.95	1.638
Over 65	4	4.25	1.500

Table 8.14 Self-assessment by education

Education	N	Mean	Std Deviation
Higher (university)	1045	3.93	1.393
Higher (non-university)	295	3.54	1.362
Secondary (general)	426	3.55	1.225
Secondary (vocational)	175	2.95	1.355
Primary	90	2.89	1.146
Other	45	3.69	1.474

Length of time learning English

Unsurprisingly, the longer people have been learning the language, the higher they rate themselves, on the whole (Table 8.15).

Nor will it be a surprise to learn that those who had only been learning for 1–2 years rate themselves predominantly at either A1 or A2 (30 per cent and 37 per cent respectively). Conversely those who have been learning for more than 10 years rate themselves at the upper levels, with almost 28 per cent believing that they are C2. Such patterns are entirely as expected, which can be interpreted as some evidence of the validity

Table 8.15 Self-assessment by length of time learning English

How long	N	Mean	Std Deviation
Less than a year	62	2.85	1.401
1–2 years	113	2.29	1.280
3–4 years	218	3.01	1.307
5–6 years	427	3.42	1.218
7–8 years	415	3.67	1.197
9–10 years	410	3.87	1.256
More than 10 years	431	4.52	1.297

Table 8.16 Self-assessment by frequency of use of English

Frequency of use	N	Mean	Std Deviation
(Almost) every day	522	4.46	1.259
Once or twice a week	907	3.51	1.332
Once every two weeks	176	3.44	1.236
Once a month or less often	264	3.08	1.179
Cannot say	190	3.25	1.418
Total	2059	3.67	1.381

of the self-assessments (although an alternative interpretation could be that if respondents have been learning for so long, they may feel that they ought to be at the higher levels, even if they are not!).

Frequency of use of English

Those who report using the language almost every day rate themselves substantially higher than those using the language less frequently (Table 8.16).

Indeed, 52 per cent of those who claim they use the language every day rate themselves at C1 or C2. Yet 60 per cent of those who use English only once a month or less nevertheless rate themselves at either B1 or B2. Of those reporting such infrequent use of English, however, 83 per cent report learning the language for at least five years.

Self-assessment by mother tongue and English proficiency

It is commonly held that learners from different cultures self-assess in different ways, such that some cultures are held to over-estimate their abilities and other cultures are said to be more modest, under-estimating their language levels. Whilst we do not have data on respondents' cultures, we do have information on their mother tongue and so it is possible indirectly to explore this issue.

Table 8.17 reports correlations between self-assessment and test scores, by mother tongue and skill tested.

Portuguese- and Italian-speaking learners show the highest correlation between self-assessment and test score, although the size of the sample is small, whereas Dutch and Swedish speakers show a low or no relationship between self-assessment and test scores. However, this may be due to differences in language proficiency. Table 8.18 reports the mean self-rating, by mother tongue. The size of the sample will affect the correlations, and so these data must be interpreted with caution.

Icelandic speakers rate themselves highest, followed by Danish speakers, then German and Norwegian speakers. Discounting Irish and

Table 8.17 Correlations of CEFR level (Overall self-assessment) with calibrated ability on skill tests

Mother tongue	Reading	Writing	Listening
Danish	.605	.565	.461
N	37	42	36
Dutch	.374	.338	.290
N	83	104	59
Finnish	.652	.635	.434
N	128	121	98
French	.606	.681	.646
N	56	60	44
German	.432	.532	.305
N	149	151	148
Icelandic	.568	.505	NS
N	17	33	21
Italian	NS	.836	.822
N	8	10	9
Norwegian	.497	.580	NS
N	34	36	31
Portuguese	.801	NS	.776
N	15	11	14
Spanish	.468	.617	.622
N	61	41	34
Swedish	NS	NS	NS
N	15	14	12

Table 8.18 Mean Overall CEFR rating, by mother tongue

Mother tongue	N	Mean	Std Deviation
Icelandic	71	4.37	1.233
Danish	116	4.27	1.379
German	450	4.09	1.303
Norwegian	103	3.94	1.290
Swedish	43	3.77	1.411
Italian	27	3.74	1.457
Portuguese	40	3.58	1.300
Greek	11	3.45	1.508
Dutch	252	3.44	1.247
Finnish	349	3.44	1.369
Spanish	137	3.39	1.274
French	160	3.13	1.273
Irish	2	2.00	.000

Greek speakers, because of small numbers, French and Spanish speakers rate themselves lowest.

However, in order to understand better the relationship between mother tongue and self-assessed language ability, it is necessary to control for language ability.

Three separate analyses were conducted: for reading ability, for writing ability and for listening ability. The results should be interpreted with some caution, paying attention to n size. It is probably best to concentrate on the results for Danish, Dutch, Finnish, French, German, Spanish and Other L1. Sample sizes for Greek, Irish, Portuguese, Icelandic, Swedish and even Italian and Norwegian are too small to be worthy of consideration.

1) Reading ability

In the Reading study, the independent variable was mother tongue, and the dependent variable was the informants' self-assessment of their reading ability based on 18 Detailed SA statements. Participants' calibrated scores on the Reading test were used as the covariate in this analysis.

After adjusting for reading ability, there was a significant difference across mother tongues on self-assessed reading ability, $F\ (13,609) = 3.472$, $p = .000$, eta squared $= .069$.

The means adjusted for reading ability are in Table 8.19.

Ignoring Swedish and Portuguese speakers because of the small sample size, the highest self-assessments of reading are given by Danish speakers, followed by speakers of German. However, speakers of languages not in DIALANG – typically the languages spoken by immigrants – rate

Table 8.19 Mean Detailed self-assessments, adjusted for reading ability

Mother tongue	Mean	Sample size
Danish	.917	37
Swedish	.853	12
Portuguese	.836	10
German	.833	135
Norwegian	.824	24
Other L1	.802	98
Spanish	.766	54
Icelandic	.746	13
French	.676	41
Greek	.662	5
Dutch	.542	83
Finnish	.483	103
Italian	.362	7
Irish	−.262	2

Table 8.20 Mean Detailed self-assessments, adjusted for writing ability

Mother tongue	Mean	Sample size
Italian	1.299	10
Spanish	1.069	37
Danish	.987	38
Portuguese	.964	9
German	.949	138
Other L1	.849	98
French	.755	54
Swedish	.734	9
Finnish	.722	94
Norwegian	.641	29
Icelandic	.568	23
Dutch	.549	101
Greek	.541	5

themselves comparatively highly, higher even than speakers of languages related to English like Icelandic and Dutch.

2) Writing ability

After adjusting for writing ability, there was a significant difference across mother tongues on self-assessed writing ability, $F\ (12,631) = 3.299$, $p = .000$, eta squared $= .059$.

The means adjusted for writing ability are in Table 8.20.

Again ignoring small sample sizes, Danish and German speakers, as well as speakers of other languages, rate themselves highly, whereas Dutch and Finnish speakers rate themselves low. Unusually, however, Spanish speakers also rate themselves highly in writing ability.

3) Listening ability

After adjusting for listening ability, there was a significant difference across mother tongues on self-assessed listening ability, $F\ (11,521) = 4.633$, $p = .000$, eta squared $= .089$.

The means adjusted for listening ability are in Table 8.21.

Ignoring small sample sizes (although it is interesting to see Icelandic speakers assessing themselves high in listening ability), the highest self-assessments are again given by Danish and German speakers, but speakers of other languages rate themselves somewhat lower in listening ability. Speakers of French and Spanish rate themselves notably low.

Clearly there are differences in self-assessed language ability by mother tongue, once test-based language ability has been taken into account. Whether this is a cultural effect, or some other variable which has not been measured, is impossible to tell. Nevertheless, Danish and

Table 8.21 Mean Detailed self-assessments, adjusted for listening ability

Mother tongue	Mean	Sample size
Icelandic	.928	18
Danish	.810	35
German	.798	136
Norwegian	.798	23
Swedish	.784	10
Other L1	.651	94
Finnish	.589	77
Dutch	.566	59
Portuguese	.553	8
Italian	.426	8
Spanish	.379	32
French	.302	34

German speakers seem to assess themselves notably higher than Dutch, French and Finnish speakers. There are, however, variations by self-assessed skill. In Reading, Other L1 and Spanish speakers rate themselves somewhat lower than Danish and German speakers do, whereas in Writing, Spanish speakers assess themselves higher than any other group. Conversely, in Listening, Spanish and French speakers assess themselves notably lower than any other group. Given the variation in self-assessment for French and Spanish speakers, it seems unlikely that the nature of the self-assessment by speakers of Spanish or French can be due to cultural, or even linguistic, factors. Whether the consistently high self-assessed ability by Danish and German speakers is a cultural factor, or simply due to the linguistic similarity between those two languages and English, is hard to say. Given the fact that Dutch is equally closely related to English, however, it may well be the case that cultural factors have a role in the modest self-assessment of Dutch speakers.

The language of the self-assessment statements, and the level assessed

Users could choose to self-assess their abilities in any one of the 14 languages of DIALANG. The relationship between one's mother tongue and the language in which users self-assessed themselves is seen in Table 8.22.

It can clearly be seen that for most mother tongues, the SAs were completed in the L1. But among those whose mother tongue was not one of the 14 DIALANG languages, users overwhelmingly chose English, the language being tested, as the interface language. In addition, others choosing to take the self-assessments in English, when taking an

Table 8.22 The relationship between mother tongue and language chosen for self-assessment, for English test-takers

Mother tongue	Interface language														Total
	Danish	Dutch	English	Finnish	French	German	Greek	Icelandic	Irish	Italian	Norwegian	Portuguese	Spanish	Swedish	
Danish	92	0	20	0	0	2	0	1	0	0	0	0	0	0	115
Dutch	0	204	40	0	1	1	0	0	0	0	0	0	0	0	246
Finnish	0	0	15	331	0	0	0	0	0	0	0	1	0	0	347
French	0	0	10	0	150	0	0	0	0	0	0	0	0	0	160
German	3	0	80	0	3	362	0	0	0	0	0	0	0	0	448
Greek	0	0	4	0	1	1	5	0	0	0	0	0	0	0	11
Icelandic	0	0	6	0	0	0	0	65	0	0	0	0	0	0	71
Irish	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2
Italian	0	0	9	0	0	2	0	0	0	16	0	0	0	0	27
Other L1	5	5	227	2	23	34	0	0	0	0	14	0	4	0	314
Norwegian	1	0	2	0	0	0	0	0	0	0	98	0	0	0	101
Portuguese	0	0	10	0	1	0	0	0	0	0	0	29	0	0	40
Spanish	0	0	31	0	0	3	0	0	0	0	0	1	101	0	136
Swedish	0	0	2	7	0	1	0	0	0	0	0	0	0	31	41
Total	101	209	456	340	179	406	5	66	2	16	112	31	105	31	2059

English test, included 80 German speakers, 40 Dutch speakers, 31 Spanish speakers, 20 Danish speakers and a few speakers of all the other languages except Irish.

Table 8.23 shows the average self-assessed Overall ability (regardless of skill) by interface language.

Ignoring those who chose the Greek interface, because of the small n size, the highest self-assessments were arrived at in Icelandic, Danish, German, Norwegian and Swedish SAs, and the lowest self-assessments were completed in Portuguese, Italian, Spanish and French (ignoring Irish because of its small n size).

Interestingly, when we correlate the mean self-assessed level by mother tongue of respondent with the mean self-assessed level by the interface language (i.e. columns two and three in Table 8.24), we get a rather low rank-order correlation of .773. Of course, the individuals are different, as shown in Table 8.22. But if there were no effect of language of self-assessment, we would expect a higher correlation, presumably approaching 1.00. However, a paired t-test showed no significant difference in mean scores ($t = -.341$, $p = .739$). Nevertheless, this result suggests that in future research attention needs to be paid to the language in which the self-assessments are made.

Table 8.23 Mean overall SA by interface language

Interface language	N	Mean	Std Deviation
Danish	101	4.31	1.263
Dutch	209	3.41	1.132
Finnish	340	3.39	1.351
French	179	3.16	1.327
German	406	4.08	1.234
Greek	5	4.80	1.095
Icelandic	66	4.33	1.244
Irish	2	2.00	.000
Italian	16	3.31	1.401
Norwegian	112	3.96	1.269
Portuguese	31	3.35	1.112
Spanish	105	3.25	1.199
Swedish	31	3.77	1.334

Table 8.24 Mean self-assessed level by mother tongue, or by interface language

Language	Mother tongue	Interface language
Danish	4.27	4.31
Dutch	3.44	3.41
Finnish	3.44	3.39
French	3.13	3.16
German	4.09	4.08
Greek	3.45	4.80
Icelandic	4.37	4.33
Irish	2.00	2.00
Italian	3.74	3.31
Norwegian	3.94	3.96
Portuguese	3.58	3.35
Spanish	3.39	3.25
Swedish	3.77	3.77

This chapter has shown the value of a measure of self-assessed language ability in predicting performance on language tests. It has also shown some variation in self-assessment by a number of background variables, including 'culture', as measured by mother tongue, and further research in this area is clearly called for. We have also seen the value of self-assessment as a pre-estimator of language ability, in order to decide which test to give users.

However, the value of self-assessment is not merely a matter of empirical investigation, at least as undertaken in this chapter. Self-assessment is an article of faith for DIALANG, since it is at the heart of

the DIALANG philosophy that an ability to assess one's language proficiency accurately will not only reveal, or lead to, greater self-awareness, it is also believed to contribute to the setting of more realistic goals for (further) language learning. This in turn will contribute to the development of learner autonomy.

Crucial in the process of awareness-raising, which is intended to get learners to reflect critically on their language proficiency, to self-diagnose, as it were, is a comparison of test results with self-assessments. The aim is not simply to show learners that there is a potential mismatch between their self-assessment and the test results, and certainly not to suggest that the self-assessment is inaccurate. Rather it is to call learners' attention to any discrepancies, to suggest a number of possible reasons why such a mismatch might occur and to encourage learners to reflect on the meaning and implications of such possibilities. That is the aim of the explanatory feedback presented in the feedback section of DIALANG. Simple quantitative analyses of the relationship between self-assessment and test results, whilst interesting in their own right, do not result in explanations of that relationship. We will explore users' reactions to the explanatory feedback and to any discrepancies between their self-assessment and their test results in Chapter 14, and I will comment further in that chapter on the value and limitations of self-assessment in the diagnosis of foreign language proficiency.

Chapter 9: Reading

In this chapter, we will first consider the construct of reading: what is reading and what reading is being tested in DIALANG? Then we will look at the DIALANG reading test specifications, and describe what the test items are intended to measure. We will then consider how an ability to read in a foreign language might relate to various reader variables, and what evidence there is for the development of such an ability. Finally, I will ask the question: what do we still need to know and explore in order to understand how best to diagnose strengths and weaknesses in reading ability?

The construct

Reading is a highly complex activity and it has been the object of a wealth of research. Many different theories of reading ability have been developed, especially for reading in one's first language, and, as we saw in Chapter 2, numerous attempts have been made to diagnose aspects of reading behaviour in L1.

Although there is much less research in a foreign language, it is nevertheless the case that the complexity of reading in a foreign language is equally well established. Alderson (2000) gives a comprehensive account of the various theories that have been developed, and of the research into the variables that can affect foreign language reading ability. It is commonplace to distinguish between the process of reading and the product – between how one reads and what one understands as a result of that reading.

Approaches to explain the process include traditional views of reading proceeding from letter to word identification, from word to sentence meaning and from understanding sentences to understanding whole texts. Such bottom-up approaches, as they are called, are often contrasted with top-down approaches, where readers are believed to approach texts with a series of expectations, based upon their knowledge of the world, of what the text will be about. Top-down readers are held

to use their expectations to hypothesize – to predict – text meaning, on the basis of which they predict the meaning of sentences, and words within them. According to this approach, readers may even not need to process the constituent letters or symbols in text. Other approaches characterize reading as being interactive, between the print on the page and what is in the reader's head, such that the process is both top-down and bottom-up. Meaning is thus seen as resulting from an interaction between the reader, the text and the task a reader assigns him- or herself – his or her reading purpose.

Similarly, the product of reading is seen to vary widely across readers and readings. One may understand the literal meaning of a text, but not understand its implicit meanings. One may understand the gist of a text, but not its author's intentions. One may be able to identify explicit facts in a text, yet not be able to separate important facts from supporting and less relevant detail. Variations in understanding are, in theories of reading, often associated with so-called skills: understanding the main idea of a text is thus seen as both a product of reading and as a subskill of reading ability.

Clearly many variables can affect both the process and the product of reading. These variables may relate to the text being read – the language in which it is written and the problems it presents to the reader, the concepts it contains, the organization of those concepts or the clarity with which they are expressed, and even variables like text layout and font can affect what one understands. Similarly reader variables, such as the readers' purpose, their motivation, their command of the language, their cognitive development, their knowledge of the world, and their willingness to make sense of the text will all have an effect on both process and product.

Relevant especially in a testing context are the tasks that a reader might decide to engage in, or be required to perform when reading – do they take notes or attempt to summarize the text, are they merely looking to locate a particular fact, or are they trying to understand the author's intention in writing the text? Such task variables and more will affect the outcome of reading, and it is above all with outcomes, with products, that assessment is concerned. However, diagnostic testing must also be concerned with understanding how readers have reached their understandings, and so insight into the reading process is also important.

There are many aspects of reading that are probably beyond testing and assessment. It is very difficult to assess meaningfully whether somebody has read a novel with great pleasure, or understood the subtleties of religious tracts. The very act of assessing and testing will inevitably affect the reading process, and the fact that a learner has answered a question posed by a tester incorrectly does not necessarily mean that he or she has not understood the text in other ways or to his or her own satisfaction.

Thus there are many limitations on what can be measured, and in the DIALANG system there are also limitations on how it can be measured. The reader's response is not evaluated in DIALANG by human beings: the computer does that. And texts are displayed on computer screens, rather than in print, which necessarily limits the sort of texts that can be tested, and the sorts of processes that can be assessed. We are also limited in our knowledge about what develops as a foreign language user becomes more proficient as a reader of foreign language texts: research in this area is in its infancy. Thus the constructs of reading that can be measured in tests like DIALANG are bound to be more limited than the multitude of possible aspects of reading that theories suggest could or indeed should be measured. Such limitations should not, however, lead us to despair: meaningful assessment of relevant aspects of the reading construct is still possible, as we shall show in this chapter.

What does DIALANG measure?

The DIALANG Assessment Specifications (DAS) describe in some detail the construct that the reading tests are intended to measure. These specifications are in part derived from the Common European Framework (CEFR), but the CEFR does not contain any description of a reading construct or an underlying theory (see Alderson *et al.*, 2004, for a discussion of the weaknesses of the CEFR in this regard). For example, the CEFR virtually dismisses any psycholinguistic approach to understanding foreign language proficiency, and does not consider, much less describe, the sorts of mental operations that a reader might undertake when reading. This is unfortunate, since the diagnosis of reading ability, as we have seen in Chapter 2, surely involves some notion of what it means to be able to read. DIALANG therefore had to have recourse to other sources for insight into what sort of operations or processes might be involved in reading that could provide useful diagnostic information.

In addition, although the CEFR describes in some (unsystematic) detail the sorts of domains and text types that a language user might be expected to understand, no indication is given in the CEFR as to which of these might apply to which level (A1 to C2). Given the lack of a theory of the development of foreign language reading ability (not only in the CEFR but in applied linguistics more generally), and given that there is not even much empirical evidence regarding how foreign language readers develop, DIALANG recognized the impossibility of developing specifications for each CEFR level separately.

Instead, the DAS merely indicated to item writers the sorts of things they might take into consideration when writing test items aimed at the various CEFR levels. As we saw in Chapter 5, item writers were asked to indicate what level was intended by each item, but this did not necessarily bear any close relationship to the empirical difficulties that resulted from the piloting.

Table 9.1 DIALANG Assessment Specifications for Reading

Cognitive Processing / Dominant intention/ Purpose	Comprehending		Comprehending + transforming or restructuring knowledge/creating connections		Comprehending + reasoning/inferencing/ interpreting/inventing/generating/ discovering	
	Events	Facts	Events	Visual images, facts, mental states, ideas	Ideas, mental states, alternative worlds	
To locate information (functional)	Timetables TV/radio guides X This week guides	Environmental print: signs, adverts Telephone directories Figures	E.g. make an itinerary using several information sources	Utilize lists of contents Read/Check minutes for specific points		
To acquire new information (referential/efferent)	News, travelogue, report of activities, recipes, instructions, directions, biographies	Descriptions, definitions	News, narrative report, instruction, telegram, announcement, circular, summary of activities	Directions, description, technical description, science report/summary	Expository texts, academic essays/ article, book review, commentary	The traditional literary genres and modes can be placed under one or more of these four purposes
To learn, to extend one's world view, to cultivate the mind (reflective)			Popular science articles, professional journal articles, popularized 'how-to books'			
To analyse/judge/ assess/evaluate/ improve text (critical)			Critical articles and reports State-of-the-art reviews		Argumentative/ persuasive texts, editorials, critical essays/articles	
To relax, enjoy vicarious experiences, to enjoy language (aesthetic, recreational)	Rhymes, jokes, anecdotes, popular magazines		Occasional poetry		Causeries, columns	

Adapted from a figure developed by Anneli Vähäpassi for the IEA International Study of Writing. (Gorman *et al.*, 1988)

Nevertheless, DIALANG was quite specific in stating that it was crucial for the success of diagnosis that all test items should be piloted, their difficulties calibrated and some means be found to relate the items to the CEFR *post-hoc*. This was accomplished, as we saw in Chapter 6, through standard-setting procedures.

On the domain of reading

The Introduction to the DAS for reading states:

> *The test items cover three aspects of reading: understanding the main idea, making inferences, and understanding specific details. The task types used are multiple-choice questions, short-answer questions and two kinds of gap-filling tasks.* (p. 6)

Table 9.1 provides an overall orientation to the domain of reading comprehension.

DIALANG, like many assessment systems, focuses on reading that aims primarily at retrieving *information* from different forms and types of texts. These are most closely represented by the first two rows of the table. Other kinds of reading also appear in the DIALANG task pools – such as reading entertaining and argumentative texts – but they are covered less comprehensively than various kinds of informative texts.

The cells of the table contain various text types/forms. The classification is, of course, not simple and clear-cut, i.e. one text may have elements of different forms. Neither does this tabular layout necessarily reflect the 'difficulty' of various text types/forms. Rather, the table illustrates how the various dimensions of texts might interlink.

Reading comprehension is basically an internal cognitive activity, but in order to demonstrate to others that comprehension has occurred, some external activity is required. The scope and nature of such external activity may vary considerably and, in reference to Table 9.1, there would normally be progression from left to right. In case of the rightmost column, reading comprehension might be displayed by comments written on the margins of a text or by producing more extended comments on the text including various kinds of syntheses and formal reviews. This type of reading-cum-writing is beyond the scope of DIALANG.

Text forms

There are different ways of classifying text forms, but one widely used classification is that provided by Werlich (1976, 1988), and this is the main source for the DIALANG approach. The main categories of the framework shown in Table 9.2 (descriptive, narrative, expository, argumentative and instructive) were considered in the selection of texts for the reading tasks.

Table 9.2 Text types in DIALANG

Text forms		**Examples (text types)**
Descriptive	• impressionistic descriptions	e.g. travel accounts
	• technical descriptions	e.g. reference books
Narrative	• stories, jokes	
	• reports: biographical notes, news, historical accounts	
Expository	• definitions	brief, one-line dictionary definitions
	• explications	broader accounts of (especially) abstract phenomena, e.g. newspaper articles, educational materials
	• outlines	e.g. initial abstract, introductory paragraph
	• summaries	. . . of phenomena, e.g. in a thesaurus
	• text interpretations	e.g. book review
Argumentative	• comments	e.g. newspaper leader, letter-to-the-editor, column, book/film review . . .
	• formal argumentation	e.g. scientific articles
Instructive	• personal instructions	e.g. signs, instructive advertisements
	• practical instructions	e.g. recipes, technical instructions
	• statutory instructions	e.g. directions, rules, regulations, law text

Writer's point of view

Another aspect to be considered when choosing texts is the **writer's point of view**. Is the writer expressing

- a fact (objective)

or

- an opinion, attitude, mood, wish (subjective)?

Test development teams were required to ensure that both objective and subjective views are represented in the selection of the stimulus texts.

Difficulty of the reading tasks

The **difficulty** of the task depends on a variety of factors related to the texts, the tasks, the reader, and their interaction. See Table 9.3.

Table 9.3 Factors affecting text difficulty

Text:	text contents: the level of abstraction, information density, theme, text form and type, contextualization, cultural conventions writers' styles: their use of vocabulary and structures, the way they build their text (cohesion, coherence), the use of redundancy
Reader:	shared background, language skills, strategies, other personal characteristics
Questions/tasks:	The tasks/items should span the whole range of difficulty. Both easy (even for clients of lower levels of ability) and demanding tasks (even for highly proficient clients) should be included in the item pool. The same text may be used as a source text for an easy item and for a more demanding item (easier to do with single-item tasks). Questions asking for specific facts are usually easier than questions requiring synthesis, analysis or inference.

In constructing assessment tasks, the test development teams paid attention to factors which contribute to text difficulty.

Selection of materials

In the construction of assessment instruments the following points were taken into consideration:

- Topics should not be too general (to avoid the testing of world knowledge) but not too specific either; they may be contemporary, but not too specifically bound to a certain context or event to avoid the texts ageing too quickly.
- Materials used should be copyright-free or able to be used without copyright restrictions for educational purposes.

Useful sources were selected, such as copyright-free newspapers and magazines, advertisements, signs, travel brochures and reference books.

Skills tested

To assess reading comprehension, items measure the user's ability:

1 to understand/identify the main idea(s), main information in or main purpose of text(s);

2 to find specific details or specific information;
3 to make inferences on the basis of the text by going beyond the literal meaning of the text or by inferring the approximate meaning of unfamiliar words.

When the comprehension of a text's main idea/information/purpose is to be assessed (reading skill 1), the answer does not depend on understanding the meaning of a single word/concept. However, if this word/concept is the keyword/key concept in the text, a question may focus on it, if either the meaning of the concept can be inferred from the context, or the meaning of the concept can be assumed to be known on the basis of general (adult) knowledge.

Sample test items

The screenshots in Figures 9.1, 9.2, and 9.3 are taken from the English Reading test in Version 1, and show how the items are presented on the computer screen (here displayed via the Review Tool).

Analysis of the results of the piloting of DIALANG reading tests

As Chapter 5 demonstrated, piloting was fairly successful for English, and so we will concentrate here on describing the results for the English Reading items. When sufficient numbers of users have taken the DIALANG pilots for the other languages, it will be possible to compare these results for those of up to 13 other languages.

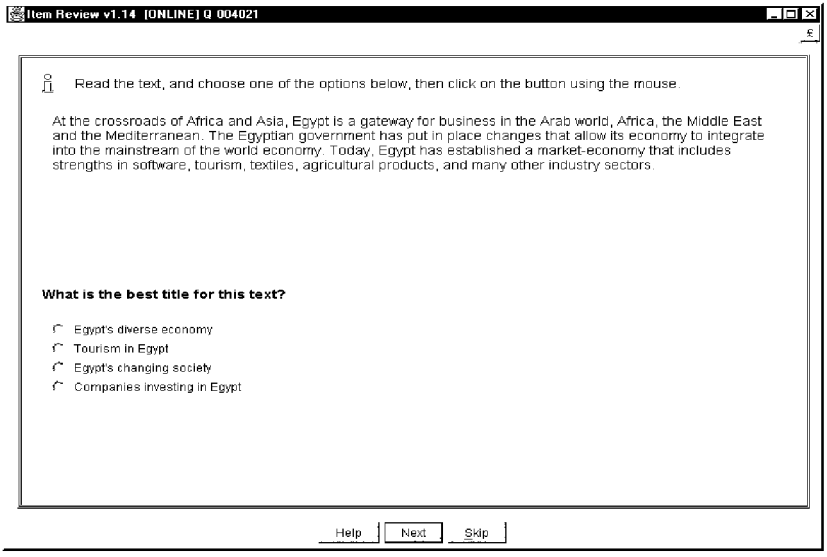

Figure 9.1 Main idea

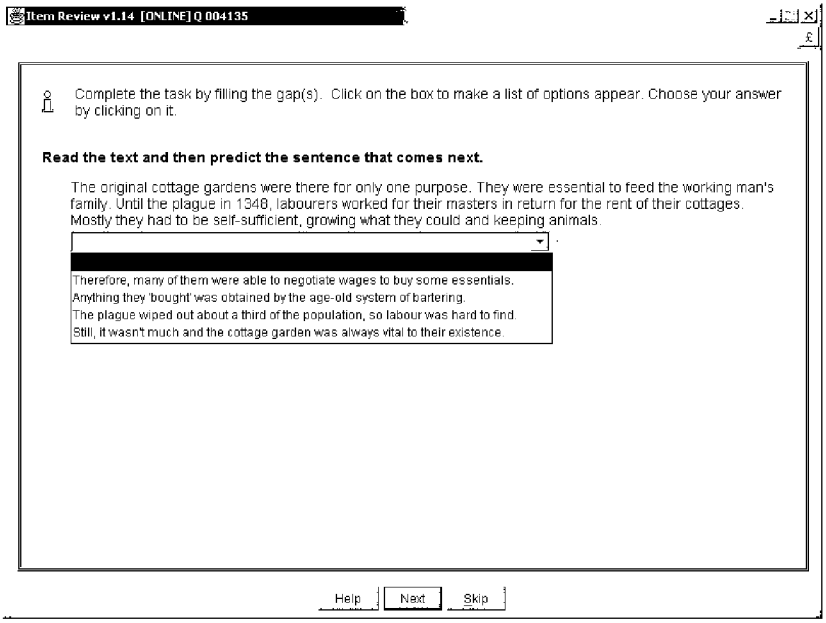

Figure 9.2 Inference

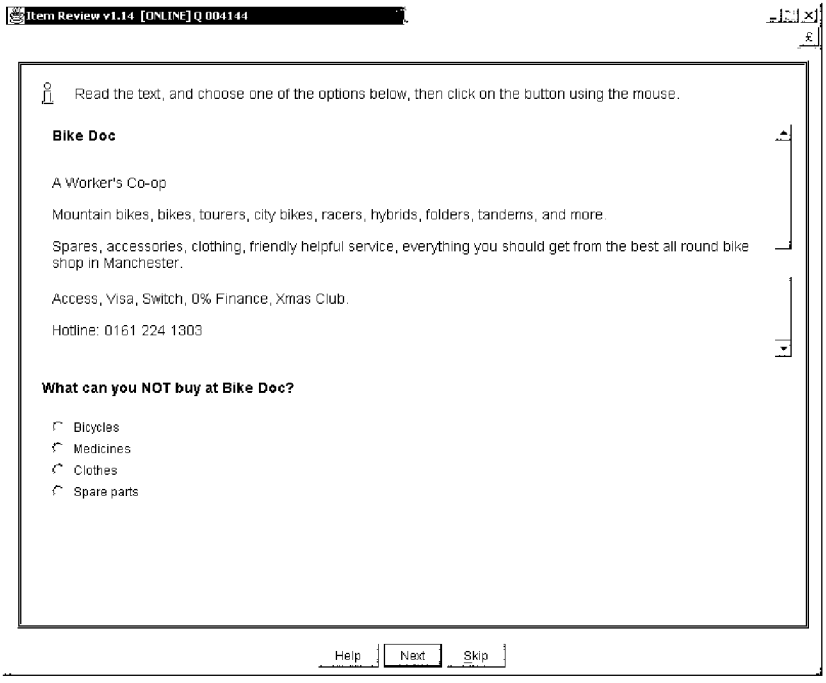

Figure 9.3 Specific detail

Table 5.11 showed that, in fact, more Reading items had to be rejected after piloting than any other skill (10 out of a total of 60 tems piloted, or 17 per cent). One problem was that for several short-answer or text-entry items, there were simply too many acceptable responses. In addition, two items, intended to be at C2, had no clearly acceptable answer.

It is quite interesting to note which subskills proved the most problematic. Thirty-three per cent of items testing 'reading intensively for specific detail' had to be rejected, but only 17 per cent of items testing 'inferencing' were rejected, and 6 per cent of items testing 'identifying main idea'. This is unusual since it is normally considered that items testing ideas that are not explicitly contained in the text, i.e. inferencing items, are harder to write than items testing explicit information because there is likely to be less agreement among those reviewing the items as to what is a legitimate inference.

Table 9.4 shows the numbers of items testing each subskill in Version 1.

An important question is to what extent items testing one subskill are more difficult, or easier, than items testing other subskills. The DIALANG Assessment Specifications make it clear that no difference in difficulty is intended across subskills. Table 9.5 reports a one-way analysis of variance of those items now in Version 1.

Although the ANOVA results just reach significance, *post-hoc* contrasts show no significant differences across pairs. This confirms research reported in the literature on foreign language reading (Alderson, 2000) that there is no reason to believe that any one of these subskills should be inherently more difficult than any other subskill.

Since each item has also been standard-set to a CEFR level, it is possible to see to what extent the different subskills correspond to different CEFR levels – is it the case that, for example, inferencing items are more prevalent at C2 than at A2? If we crosstab CEFR level by subskills, we find no significant association ($\chi^2 = 6.844$, $p = .336$).

Table 9.6 confirms that an ability to respond correctly to, for example, a main-idea item is not in and of itself necessarily indicative of a particular CEFR level. This has implications for a theory of the development of reading ability, as we shall discuss later.

Table 9.4 Number of items per subskill in Version 1

Reading subskill	Number of items	Percent of total
Inferencing (including lexical inferencing)	25	50.0
Identifying main idea/Distinguishing from supporting detail	17	34.0
Reading intensively for specific detail	8	16.0

Table 9.5 Difficulty by subskill

Subskill	N	Mean	Min	Max	SD		
Identifying main idea/ Distinguishing from supporting detail	17	−.071	−.84	.35	.3025		
Inferencing (including lexical inferencing)	25	.128	−.83	1.23	.4662		
Reading intensively for specific detail	8	−.250	−.57	.04	.2032	F = .3.355	P = .043

Table 9.6 Subskills tested by CEFR level

Subskill	A1	A2	B1	B2	Total
Main idea	1	3	10	3	17
Inferencing	2	2	11	10	25
Specific detail	1	2	5		8
Total	4	7	26	13	50

Of the items that survived into Version 1, only one is a text-entry type and there are no short-answer questions, as a result of the piloting. There are 39 multiple-choice items and 10 drop-down items (which are in any case a form of multiple-choice). This bias towards multiple-choice is perhaps inevitable in a computer-marked test, given the problem of indeterminacy of inferencing, especially if users are free to use their own words in response.

To see whether one item type was more difficult than another, a one-way ANOVA was run, as shown in Table 9.7. The means are significantly different, with drop-down being significantly harder than multiple-choice. The one text-entry item was actually easier than one

Table 9.7 Difficulty by item type

Item type	N	Mean	Min	Max	SD		
Drop-down	10	.223	−.526	1.226	.5284		
Multiple-choice	39	−.075	−.839	.771	.3359		
Text-entry	1	.671	.671	.671		F = 4.028	p = .024

Table 9.8 Number of items at each CEFR level (Author's Best Guess) in and after the pilot

Skill/Level	A1	A2	B1	B2	C1	C2
K items in pilot	6	13	12	12	11	6
K items surviving pilot	6	7	11	12	10	4

multiple-choice item and two drop-down items. However, there is unlikely to be a theoretical explanation of such differences, which are likely to be due to test content. Nevertheless it might be worth exploring this further with a much larger database of items.

Table 9.8 shows how many items there were at each estimated level of the CEFR during and after the pilot: most attrition took place at level A2, followed by C2.

However, these CEFR levels are merely what the item writers intended the items to be at. Table 9.9 shows the calibrated difficulty levels of the piloted items, at estimated levels A, B and C. There is no significant difference among the levels, confirming the doubts one might have had of the accuracy of the author's best guess (ABG) as to the level of the items they wrote.

Table 9.9 Difficulty by Author's Best Guess

Item level	N	Mean	Min	Max	SD		
A	13	−.184	−.83	.77	.4304		
B	23	.011	−.84	1.05	.3895		
C	14	.154	−.26	1.23	.3540	F = 2.535	P = .090

Table 9.10 gives the number of items from Version 1 that were calibrated and standard-set at each CEFR level. The lack of items at C1 and C2 confirms the difficulty of producing objectively markable items at higher levels of reading ability.

Table 9.10 Level of calibrated Reading items in Version 1

Skill/Level	A1	A2	B1	B2	C1	C2
K items	4	7	26	13	0	0

Reading ability

So far we have examined the construct underlying the Reading items, and how the Reading items performed during the pilot testing. In this

Table 9.11 Descriptive statistics for the Reading test

N	Minimum	Maximum	Mean	SE mean	Standard Deviation
718	−.625	2.159	.399	.017	.464

next section we explore how learners performed on the pilot tests. Since pilot test-takers took different items depending on the booklet to which they were assigned (see Chapter 5), it was necessary to calibrate users' scores, using Item Response Theory, in order to arrive at comparable scores. The ability scores reported in this section are thus expressed in logits, where a minus indicates a low level of ability and positive figures indicate higher levels of ability.

First, in Table 9.11, we present the descriptive statistics for the Reading test as a whole, bearing in mind, as presented in Chapter 5, that these scores are derived from four different booklets, but are calculated using IRT with the anchor items.

In what follows, we present the results of learners' Reading test performances, according to the background variables on which data was gathered during piloting (Chapter 5).

Reading ability by mother tongue

A one-way ANOVA showed highly significant differences across the various mother tongues ($F = 4.745$, df 13, $p = .000$). However, because of the varying sample sizes, the only significant contrasts were between German, on the one hand, and Finnish, Spanish and Other L1s on the

Table 9.12 Reading ability by mother tongue

Mother tongue	Mean	N
Icelandic	.677	17
German	.570	149
Irish	.524	2
Norwegian	.517	34
Swedish	.456	15
Dutch	.449	83
Danish	.436	37
Italian	.420	8
Portuguese	.364	15
French	.343	56
Finnish	.333	128
Other L1	.258	108
Spanish	.202	61
Greek	.033	5

Table 9.13 Reading ability by sex

Sex	Mean	N
Female	.430	437
Male	.350	281

other, and between Icelandic, Dutch and Norwegian on the one hand, and Spanish on the other. Icelandic and German speakers performed notably better than speakers of other languages. See Table 9.12.

Reading ability by sex

If we consider the difference between males and females in reading ability (Table 9.13), we find that females were significantly better at reading than males ($p = .023$).

Table 9.14 Reading ability by age

Age	Mean	N
Under 18	.211	23
18–25	.390	450
26–35	.396	120
36–45	.444	76
46–55	.484	38
56–65	.564	11

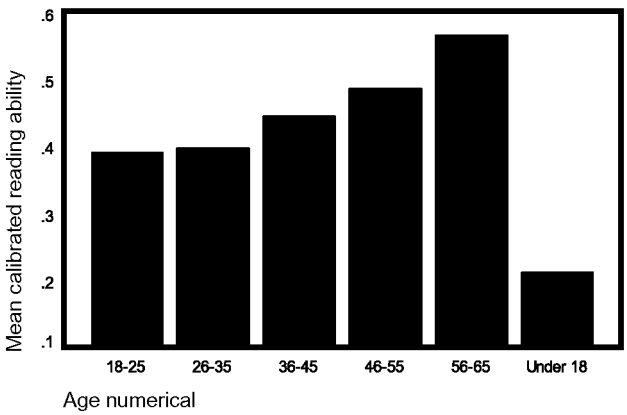

Figure 9.4 Reading ability by age

Reading ability by age

Table 9.14 presents the results of the reading test by age. A one-way ANOVA revealed no significant difference by age ($F = 1.472$, $p = .197$).

Nevertheless, Figure 9.4 shows the expected trend.

Reading ability by education level

A one-way ANOVA shows the differences between education levels to be highly significant ($F = 10.208$, $p = .000$), with University being significantly different from Secondary and Primary. Results are in Table 9.15.

Reading ability by how long users have been learning English

Table 9.16 presents reading ability by length of time learners have been learning the language. The difference among means is highly significant ($F = 25.602$, $p = .000$). Figure 9.5 shows the differences graphically, with reading ability increasing with length of time learning the language, albeit not dramatically.

Table 9.15 Reading ability by education level

Education	Mean	N
Primary	.138	29
Secondary (vocational)	.155	68
Secondary (general)	.343	153
Higher (non-university)	.360	101
Higher (university)	.495	347
Other	.554	20

Table 9.16 Reading ability by how long users have been learning English

Length of time	Mean	N
Less than a year	.199	19
1–2 years	−.142	44
3–4 years	.179	77
5–6 years	.386	159
7–8 years	.396	141
9–10 years	.476	130
More than 10 years	.649	148
Total	.399	718

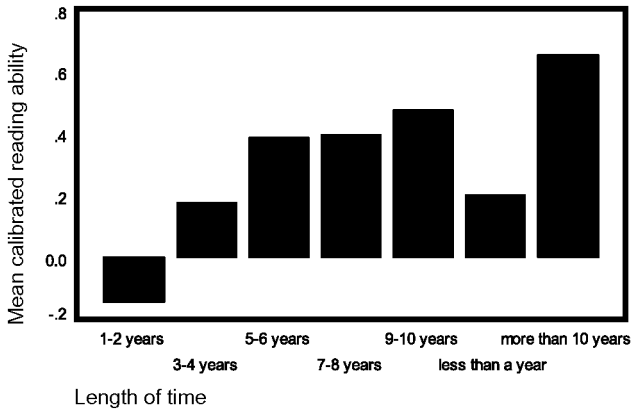

Figure 9.5 Reading ability by length of time learning English

Reading ability by frequency of use

Examining the effect of frequency of use of the language on reading ability, a one-way ANOVA showed significant differences ($F = 15.859$, $p = 0.00$) with 'almost every day' being significantly higher than other frequencies and trends in the expected direction, as Table 9.17 shows.

Table 9.17 Reading ability by frequency of use

Frequency of use	Mean	N
Cannot say	.311	70
Once a month or less often	.249	94
Once every two weeks	.374	64
Once or twice a week	.339	308
(Almost) every day	.619	182
Total	.399	718

Reading ability by self-assessed reading

The correlation between reading ability and overall self-assessed reading was .54. Table 9.18 shows the mean reading ability by Overall self-assessment.

With the exception of A2, the trend was in the expected direction, and the differences were highly significant ($F = 58.670$, $p = .000$). Figure 9.6 presents the data graphically.

Relationship between reading subskills and macro skills

It will be remembered that pilot test-takers who were assigned a Reading Test booklet were also required to take a test of Grammar. Although

Table 9.18 Reading ability by self-assessed reading

CEFR level of Overall SA	Mean	N
A1	.086	53
A2	−.0005	80
B1	.220	151
B2	.454	256
C1	.702	94
C2	.789	84

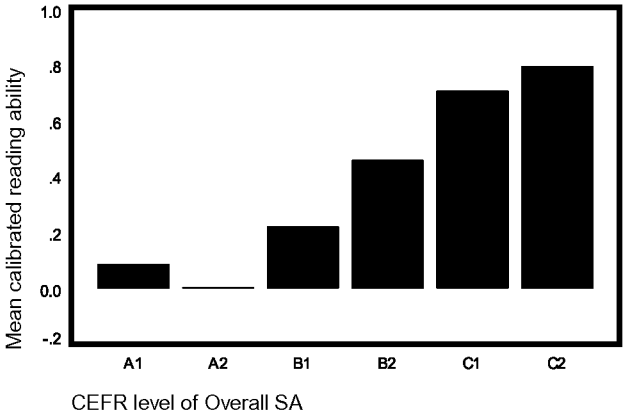

Figure 9.6 Reading ability by Overall self-assessed reading ability

we cannot compare performances on Reading with those on Writing or Listening, we can nevertheless examine the relationship between Reading and Grammar. The correlation between the macro skills of Reading and Grammar was .685 (Spearman), with a sample size of 718. In Table 9.19 we examine how the various subskills of reading related to the macro skills.

Whilst Table 9.19 shows a moderately close relationship between the subskills of reading and reading overall, what is interesting about Table 9.20 is that it shows not insubstantial levels of correlation between reading subskills and grammatical ability, especially bearing in mind that the correlation between reading subskills and overall reading ability includes the subskill concerned and is therefore inflated.

Various factor analyses were run to see whether different factors underlie the various tests. The reading subskills were analysed in relation to overall reading ability, to overall grammatical ability and to a combination of reading and grammar. In every case, only one factor

Table 9.19 Reading subskills and the macro skill of Reading

	Identifying main idea	**Reading intensively for detail**	**Inferencing**	**Reading ability**
Identifying main idea	1	.507	.603	.756
Reading intensively for detail	.507	1	.549	.620
Inferencing	.603	.549	1	.851
Reading ability	.756	.620	.851	1

Table 9.20 Reading subskills and Grammar

	Identifying main idea	**Reading intensively for detail**	**Inferencing**	**Overall grammar ability**
Identifying main idea	1	.507	.603	.540
Reading intensively for detail	.507	1	.549	.544
Inferencing	.603	.549	1	.665
Overall grammar ability	.540	.544	.665	1

emerged, accounting for between 68 per cent and 74 per cent of the variance. The factor loadings for the subskills were roughly equal and did not differ substantially across analyses. What was interesting was that even grammatical ability had a high loading – .787 – on the reading subskills factor. Clearly there is a close relationship between grammatical ability and reading, perhaps because the grammar test items require a reading ability, or because the reading abilities tested require grammatical knowledge.

Summary and conclusion

In this chapter we have examined the construct of reading, described the specifications for the Reading tests in DIALANG and I have

presented test items based on those specifications. We have explored how well the English Reading test items performed during piloting, and we have also looked at the performance of those learners who took English pilot Reading tests.

We established that there were significant differences in reading ability, depending upon the mother tongue of the learner, with those who spoke a language more closely related to English achieving significantly higher Reading test scores. Females were significantly better at reading than males, but age proved not to be a significant variable. Educational level did prove to be significant, with university-level students performing better than those with lower levels of education. The length of time a person had been learning English was significantly related to level of reading ability, as was frequency of use of English. The use of English almost every day resulted in markedly higher levels of reading ability. Encouragingly, a learner's self-assessed level of reading showed a significant relationship to their level of reading according to the CEFR, providing evidence for the validity of the results reported by DIALANG.

The three subskills tested in DIALANG Reading tests did not, however, relate to the different CEFR levels in the sense that even low-level test items (A1, A2, B1) tested all three skills, showing that all subskills tested are relevant across the CEFR levels. Moreover, learners who achieved scores indicating they were at higher CEFR levels showed weaknesses in all three skills. It appears not to be the case that as one's reading ability develops, this is associated with an increased ability to make inferences, for example, rather than to understand the main idea.

Thus it will be important in the further development of diagnostic tests of reading to establish, first, exactly what does improve as the ability to read in a foreign language increases, and secondly, what subskills or aspects of linguistic knowledge or text processing predict such a developing ability to read. It is likely that research into the contribution of knowledge and skills related to language – vocabulary knowledge, the ability to parse sentences and clauses, the ability to follow the organization of a text, and so on, the sort of variables discussed and explored in Alderson (2000), for example – will prove to be worthy of further exploration.

Above all, it is clear that much more research is needed into how learners perform on the various DIALANG tests, and what this reveals about the development of their ability to read in a foreign language. We have not yet been able to explore, through extensive research, how learners will perform across the whole range of abilities and subskills tested in DIALANG. But hopefully the continued free availability on the Internet of DIALANG will encourage and enable researchers to use the tests to explore the constructs, to develop further hypotheses about the development of reading, and to devise new and insightful diagnostic measures, building on the foundations already established.

Chapter 10: Listening

In this chapter, we will discuss the construct of listening: what is involved in listening to and understanding a foreign language, and what listening abilities are being tested in DIALANG. Next we will examine the DIALANG Listening items, and describe what they test. We will then consider how users performed on the tests of listening in DIALANG, and discuss how the ability tested relates to various learner variables, and what evidence there is for the development of listening ability. Finally, we will ask the question: what do we still need to know and explore in order to understand how best to diagnose strengths and weaknesses in listening ability?

The construct

Like reading, only more so, listening is a highly complex ability. Curiously, however, unlike reading, the ability to understand spoken texts has received much less attention in the literature, both in listening to one's mother tongue and in listening to a foreign language. Given the additional complexity of listening over reading, this is quite surprising. Not only must a listener do all the things that a reader has to do – process language, pay attention to both language and concepts, their organization and format, relate what he or she knows to what is being heard, and so on (see Chapter 9) – but crucially, a listener normally has to do all this in real time. It is rare for somebody to have the opportunity to listen to the exact same text a second time, unless, that is, they are listening to a tape- or video-recording. Thus the processing demands of real-time listening are bound to be greater than those of the processing of written text, which can be examined again and again.

A second major difference is that the listener has to deal with sound signals which vary greatly from speaker to speaker in a way that printed text does not vary by author. Orthographic conventions are one thing, but the variability in the pronunciation of the sounds that constitute the spoken text is much greater, and listeners have to learn to cope with very varied signals. And the signals are part of a continuous stream of sound, which is not segmented in the way that many writing systems are, so

listeners have to identify the elements in the sound stream – phonemes, words, complete utterances – with little acoustic information to support them. They also have to cope with listening under often less-than-ideal circumstances – with background noise, with other speakers intervening, and often without being able to see the speaker.

In so-called interactive listening, when the listener and the speaker are both present in the interaction (even if separated by distance, such as on the telephone), at least the listener can overcome some of these problems by asking the speaker to repeat, or can indicate his or her incomprehension by facial gestures, backchannelling, and so on. Thus compensatory devices can be available to listeners which are not available to most readers. Nevertheless, in many testing situations, even this compensation is not available, which is why many listening tests present the spoken text twice to the listener, in part compensation for the lack of interactivity.

Buck (2001) gives a full account, both of theories of listening in a foreign language and of the research that has led to the identification of variables that are crucial to the understanding of spoken text. He addresses matters like the linguistic features that are peculiar to spoken texts, the importance of redundancy in the acoustic input, the nature of unplanned discourse, and deals at length with issues that make listening unique, like speech rate, phonological modification, prosodic features (intonation and stress), discourse structure, hesitation features, non-verbal signals and the differences between first language listening and second language listening. He describes the variety of skills that have been proposed as constituting a listening ability in a foreign language and he discusses a whole range of variables that have been shown to influence the way in which we can best assess a learner's listening abilities.

However, inevitably, as with other skills, the testing of listening is bound to involve compromises, in terms of the texts that learners can be asked to understand, the processes that they can be assumed to undergo, the tasks they can be given, and the sorts of inferences that can be drawn from their test-based performances. Computer-based testing is no exception in this regard, and indeed tests that are scored by computer clearly involve giving learners tasks that they may not face in real life (where it is unusual to be asked a number of questions, with four optional answers each, whilst one is listening to a spoken text). Thus it is very important, when evaluating any test of listening ability, to be aware of the limits of the possible, but also to be cognisant of the constructs that, despite the limitations, the test developers have attempted to measure.

DIALANG specifications

The DIALANG Assessment Specifications (DAS) are in part derived from the Common European Framework (CEFR), but the CEFR does not contain any description of a listening construct or an underlying

theory. Indeed, the CEFR virtually dismisses any psycholinguistic approach to understanding foreign language proficiency, and does not consider the sorts of mental operations that a listener might undertake when listening. This is unfortunate, since the diagnosis of listening ability, as we have seen above, must involve some notion of what it means to be able to understand spoken discourse. DIALANG therefore had to look to other sources for insight into what sort of operations or processes might be involved in listening that could provide useful diagnostic information.

In addition, although the CEFR describes the sorts of domains and text types that a language user might be expected to understand, the CEFR does not indicate which of these might apply to which level (A1 to C2). Given the lack of a theory of the development of foreign language listening ability (not only in the CEFR but in applied linguistics more generally), and given that there is not much empirical evidence regarding how the ability develops to understand a foreign language when spoken, DIALANG did not develop separate specifications for each CEFR level. Instead, the DAS merely indicated to item writers the sorts of things they might take into consideration when writing listening test items aimed at the different CEFR levels.

As we saw in Chapters 6 and 9, however, DIALANG insisted that all test items should be piloted and related to the CEFR, and their difficulties calibrated.

The Introduction to the DAS for listening states: '*The test items cover three aspects of listening: understanding the main idea, making inferences, and understanding specific details*'. The specifications describe how difficulty (in terms of the CEFR levels) may be varied by manipulating linguistic, psycholinguistic or sociolinguistic features of the discourse. Item writers are encouraged to sample different discourse types, from conversations or dialogues, with speakers taking turns, to more formal ones such as lectures or sermons, where speech may be primarily monologic. The discourse types vary in terms of structure, diffuseness or otherwise of information, turn-taking, participants' roles and power relations, public/private domain, topic, situational contexts and so on.

Roles of the listener and purposes of listening

The listener can act as a participant, an addressee, an audience member or a judge. In defining purposes of listening, the Council of Europe draft Framework (1996: pp. 55–56) provided useful lists under 'functional competence' in terms of macro- and microfunctions. See Table 10.1.

For the DIALANG task pool for listening, tasks based on both dialogue-type and monologue-type discourse were produced. Where monologue-type discourse is involved, a variety of discourse forms were covered. When dialogues were used as material for listening comprehension, it was generally not feasible to include very diffuse conversations, because they tend to be long and the time allocation for

Table 10.1 Purposes for listening

General purpose:	Seeking/giving factual information Finding out/expressing attitudes Recognizing persuasion/persuading Socializing (Recognizing markers of discourse structuring (e.g. opening, turn-taking, closing))

listening material was short. In the context of DIALANG, the length of listening samples is 30–60 seconds. Vignettes such as requests for advice, telephone conversations or service encounters can much more easily be incorporated into this kind of length limit, as can brief meetings between friends or acquaintances. The whole conversation does not need to be included in a listening passage; but when using an extract, care has to be taken that enough context is given in the task-specific instructions to make the extract comprehensible.

Table 10.2 Text types for listening tasks

Discourse forms		**Examples (discourse types)**
Descriptive	• impressionistic descriptions	e.g. travel accounts, sports commentaries
	• technical descriptions	e.g. presentation of a product
Narrative	• stories, jokes, anecdotes	
	• reports	e.g. news reports, features, documentaries
Expository	• definitions	brief, definitions
	• explications	broader accounts of (especially) abstract phenomena, e.g. lectures, talks
	• outlines	e.g. programme listings on the radio, timetables
	• summaries	e.g. an oral account of the plot of a book
	• summarizing	minutes of a meeting
	• interpretations	e.g. describing a book, an article, etc.
Argumentative	• comments	by any individual in any situation
	• formal argumentation	e.g. formal debate
Instructive	• personal instructions	e.g. announcements, ads, propaganda, routine commands
	• practical instructions	e.g. recipes, technical instructions, paging at an airport

The categorization of discourse forms is equivalent to that presented in the specifications for reading, but was adapted to listening. The main categories of the framework (shown in Table 10.2: descriptive, narrative, expository, argumentative and instructive) were taken into account when selecting monologue-type pieces of discourse for the listening tasks, in order to provide for enough variation in the task pool.

Speaker's point of view

Another aspect to be considered when choosing pieces of discourse for listening is the speaker's point of view and whether what is said is:

- a fact (objective)

or

- an opinion, attitude, mood or wish (subjective).

In addition to the difficulty of the spoken material, the difficulty of the listening task also depends on the nature of the task itself. Questions asking for specific facts are thought to be easier to answer than questions requiring synthesis, analysis or inference.

In the test-construction process, attention was paid to factors which contribute to the difficulty of the spoken material. Except in the case of items at the highest levels of difficulty, speakers used more or less the standard variety (or varieties) of the language and topics were kept general so that comprehension did not require a knowledge of specialist vocabulary. Detailed guidelines were given concerning authentic and scripted discourse, the quality of the speakers (combinations of distinguishable voices, male/female, old/young, etc.) and the quality and source of the recording.

Skills tested

To assess listening comprehension, items measure the learner's ability:

1 to understand/identify the main idea(s), main information or main purpose of a piece of spoken discourse (including the contribution of prosodic features);
2 to listen intensively for specific detail or specific information;
3 to make inferences on the basis of what was heard and to be able to use context to infer the approximate meaning of an unfamiliar word.

Many items contain a brief characterization of the central idea. This states where the discourse is taking place, the identity of the speaker(s), and what is happening in the situation. For instance:

- in a café, a customer returns because he has forgotten his hat/lost his wallet, and asks the waitress and a customer if they have seen it;
- on a street, a passerby sees a woman drop her scarf and calls out to her;

- on the radio, a traveller gives a description (of a restaurant in Peru, a beach in Ireland, a walk in the Alps, a boat trip in midsummer . . .);
- in a discussion, a nutritionist talks about the benefits of a diet of fish.

Test items

The item types used are multiple-choice, short-answer and gap-fill (drop-down or text-entry). When designing listening tasks, test developers had to keep two important things in mind:

1 the test-takers will hear each clip **only once**;
2 the clips must be as short as possible and a maximum of 60 seconds.

English items in Version 1

The screenshots in Figures 10.1, 10.2 and 10.3 present items as they are shown on screen in Version 1, with their accompanying tapescripts, although, of course, test-takers cannot see any tapescript.

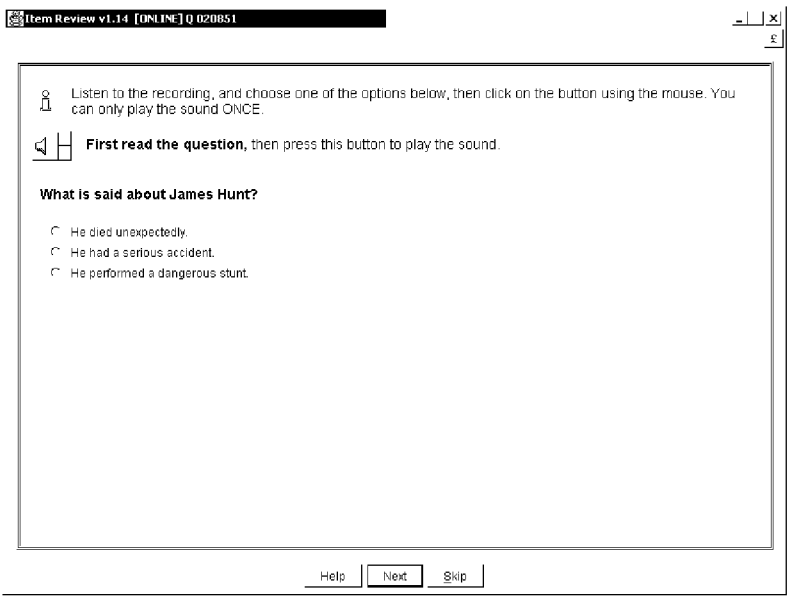

Figure 10.1 Main idea

Tapescript
James Hunt, one-time superstar of British motor racing, died of a heart attack yesterday. The death of the 45-year-old ex-World Champion stunned family and friends and plunged the sporting world into mourning.

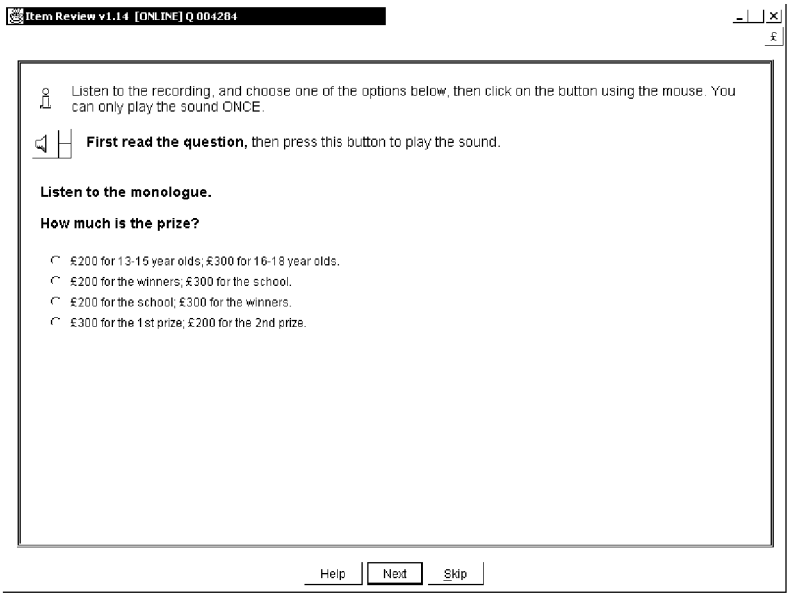

Figure 10.2 Specific detail

Tapescript
Before you leave, there is something you might want to know. There's a competition called Young Political Writer of the Year Award. And you might want to enter. Now you can enter in two different groups. There's a competition for 13 to 15-year-olds, and there's another competition for 16 to 18-year-olds. So you just put yourself in the right group for that. Now the prize is to be shared between the winner – the person who is the Best Political Writer of the Year – and the school. So the winner gets £200 and the school gets £300. Doesn't seem very fair, but that's the way it is. And also you get a trip to Parliament, to hear debates and to meet the Speaker of the House of Commons. Now to win the prize you have to write a feature article and you can choose whichever topic you want to write about. And it should be no more than 600 words long. So – get writing!

Results of piloting of English Listening

Fifty-six Listening items out of 60 piloted survived the pilot testing (i.e. 93 per cent), and of those that were rejected, two were intended to be A2, one B1 and one C1. Two were items intended to test specific details, one was intended to test main ideas and one was intended to test inferencing. Two were multiple-choice items, and two were short-answer items.

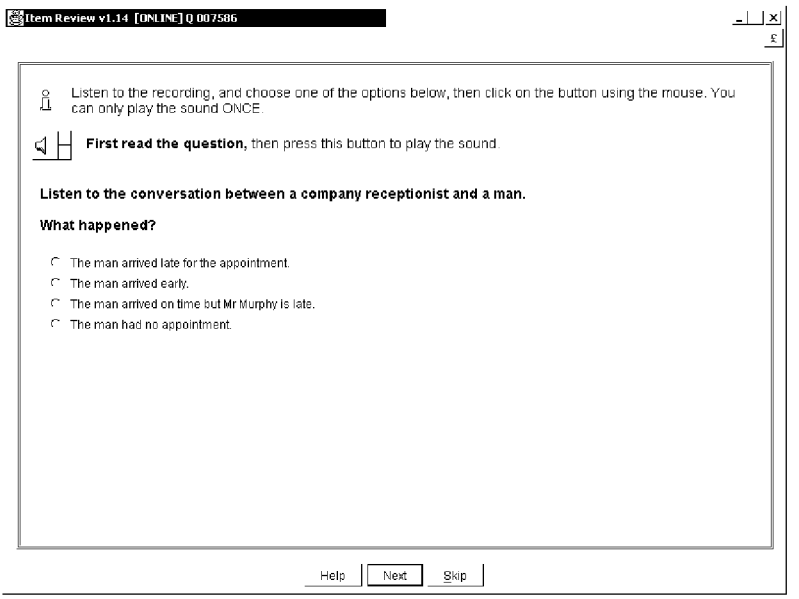

Figure 10.3 Inferencing

Tapescript
Woman: Hello. Can I help you?
Man: Yes, please. I am here to see Mr Murphy.
Woman: Take a seat. I'll get him for you . . . I'm afraid he is not in his office. Do you have an appointment?
Man: Yes, at 3 o'clock, but . . .
Woman: Ahhhh. He must have thought you weren't coming. If you'd like to wait I'll see if I can find him for you, but he may be busy . . . Or would you like to make another appointment?
Man: Oh, I'll wait a little while.
Woman: OK.

Of the surviving items there was a very high proportion of multiple-choice items (55 out of 56). The only other item was a short-answer question. Table 10.3 shows the mean logit values for these items.

Since there is only one short-answer question, and there are 11 multiple-choice items that are more difficult, there is no difference in mean scores by item type.

Table 10.4 examines the differences in empirical difficulty by estimated CEFR level (Author's Best Guess).

Although a one-way ANOVA shows a highly significant difference in mean scores, *post-hoc* tests show the only significant contrast to be

Table 10.3 Difficulty by item type

Item type	N	Mean	Min	Max	Std Deviation
Multiple-choice	55	−.004	−.466	1.418	.391
Short-answer	1	.202	.202	.202	

Table 10.4 Difficulty by Author's Best Guess

Item level	N	Mean	Min	Max	Std Deviation		
A	20	−.182	−.466	.080	.177		
B	31	.046	−.385	1.299	.390		
C	5	.444	−.137	1.418	.596	F = 6.895	P = .002

between A and C (as estimated by the authors). Moreover, if we compare ABG with empirically established CEFR levels, we find no significant relationship ($\chi^2 = -30.860$, p = .057).

In terms of subskills, there were 34 main idea items (61 per cent), 16 inferencing items, and only 6 (11 per cent) specific detail items. Table 10.5 examines differences in difficulty across subskills.

A one-way ANOVA shows no significant difference in mean scores by subskill. There is thus no reason to believe that inferencing is more difficult than understanding the main idea or that understanding explicit information is easier than the other subskills.

Table 10.5 Difficulty by subskill

Subskill	N	Mean	Min	Max	Std Deviation		
Identifying main idea/ distinguishing from supporting detail	34	−.007	−.47	1.300	.344		
Inferencing (including lexical inferencing)	16	.039	−.40	1.42	.450		
Listening intensively for specific detail	6	−.064	−.44	.88	.508	F = .163	P = .850

Table 10.6 Descriptive statistics, Listening ability scores

N	Minimum	Maximum	Mean	SE Mean	Standard Deviation
606	−.807	1.999	.47299	.02114	.520384

If we classify each subskill according to its empirically determined CEFR level and then explore the relationship between level of item and the subskills it is testing, there is no significant association ($\chi^2 = 6.998$, $p = .537$), which reinforces the conclusion that there is no relationship between an item's level or difficulty and the subskill it tests.

Listening ability

First we explore the relationship between listening ability and the various background variables on which data were gathered through the background questionnaire during piloting.

To facilitate comparisons, Table 10.6 presents the descriptive statistics for all learners taking the Listening tests (logit scores).

Listening ability by mother tongue

A one-way ANOVA showed highly significant differences ($F = 9.001$, $p = .000$). As with Reading, Icelandic and German speakers performed notably better than speakers of other languages. See Table 10.7.

Table 10.7 Listening ability by mother tongue

Mother tongue	Mean	N
Greek	1.113	1
Icelandic	.828	21
German	.669	148
Dutch	.606	59
Italian	.587	9
Swedish	.586	12
Norwegian	.560	31
Finnish	.495	98
Danish	.478	36
French	.286	44
Other L1	.187	99
Spanish	.157	34
Portuguese	.108	14

Table 10.8 Listening ability by sex

Sex	Mean	N
Female	.482	378
Male	.458	228

Table 10.9 Listening ability by age

Age	Mean	N
Under 18	.240	12
18–25	.469	370
26–35	.522	124
36–45	.515	60
46–55	.337	37
56–65	.825	2
Over 65	.383	1

Listening by sex

The difference between males and females in listening ability was not significant ($p = .577$) (Table 10.8).

Listening ability by age

Table 10.9 presents listening ability by age. The difference among ages was not significant ($F = 1.234$, $p = .287$). Nevertheless, Figure 10.4 shows the trends.

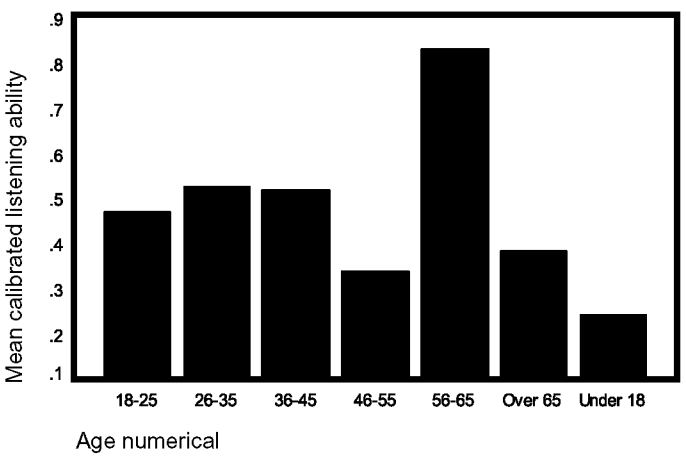

Figure 10.4 Listening ability by age

Listening ability by educational level

The differences were significant overall ($F = 4.623$, $p = .000$), although few contrasts were significant. See Table 10.10.

Table 10.10 Listening ability by educational level

Educational level	Mean	N
Higher (university)	.550	329
Secondary (general)	.456	111
Primary	.428	28
Higher (non-university)	.352	90
Secondary (vocational)	.241	38
Other	.219	10

Table 10.11 Listening ability by length of time learning English

Length of time	Mean	N
Less than a year	−.054	16
1–2 years	.043	28
3–4 years	.198	61
5–6 years	.411	127
7–8 years	.489	120
9–10 years	.629	122
More than 10 years	.656	132

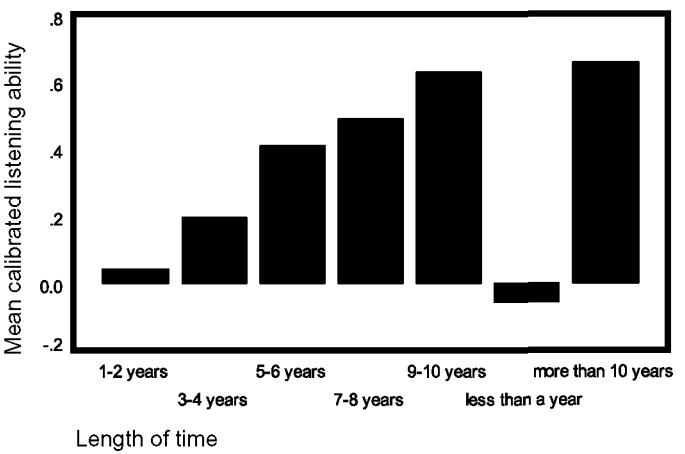

Figure 10.5 Listening ability by length of time learning English

Listening ability by length of time learning English

See Table 10.11. The differences are significant ($F = 15.612$, $p = .000$) overall and Figure 10.5 shows the expected trends graphically. Listening ability increases with length of time spent learning the language.

Listening ability by frequency of use

As expected, the differences are significant overall ($F = 12.816$, $p = .000$), with 'almost every day' being significantly higher than all other contrasts (Table 10.12).

Listening ability by self-assessed listening ability

Table 10.13 presents the data. A one-way ANOVA showed overall significance ($F = 5.643$, $p = .000$) and Figure 10.6 shows the expected trends.

Relationship between skills and subskills

As explained in Chapter 5, test-takers assigned to a Listening test were also given a test of Vocabulary. It is thus possible to compare performance on these two tests, both at macro skill and subskill levels.

Table 10.12 Listening ability by frequency of use

Frequency of use	Mean	N
Cannot say	.37131	49
Once a month or less often	.43088	85
Once every two weeks	.31791	45
Once or twice a week	.39419	278
(Almost) every day	.72430	149
Total	.47299	606

Table 10.13 Listening ability by self-assessed listening ability

SA level	Mean	N
A1	.110	58
A2	.078	55
B1	.359	164
B2	.518	135
C1	.764	104
C2	.752	90

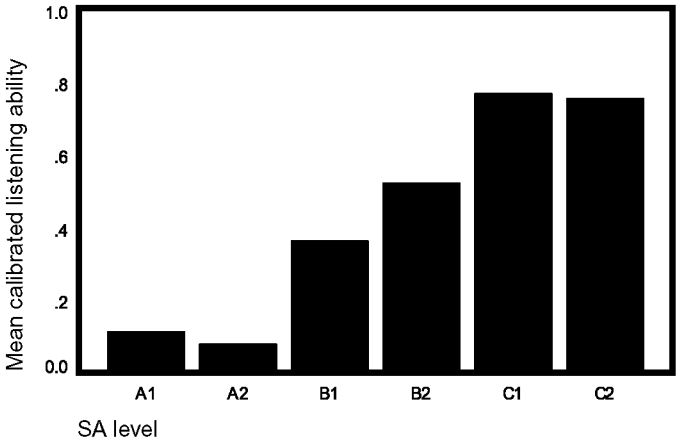

Figure 10.6 Listening ability by overall self-assessed listening ability

The correlation between Listening and Vocabulary was .65 (Spearman), sample size n = 606. Table 10.14 shows the correlations between listening subskills and the macro Listening skill, and Table 10.15 shows the relationship between listening subskills and vocabulary knowledge.

Perhaps not surprisingly, the listening subskills relate more closely to the macro skill of Listening than they do to a different macro skill, that of vocabulary knowledge. However, this is in part because the skills are correlating with themselves, as components of the macro skill. Clearly the ability to identify the main idea in spoken text and to draw inferences from such texts is related to vocabulary knowledge.

Table 10.14 Listening subskills and Listening macro skill

Subskill	Listening ability
Listening intensively for detail	.616
Listening inferencing	.813
Listening to identify main idea	.885

Table 10.15 Listening subskills and Vocabulary

Subskill	Vocabulary ability
Listening intensively for detail	.383
Listening inferencing	.560
Listening to identify main idea	.598

Table 10.16 Results of factor analysis of subskills, with Listening macro skill

	Factor 1
Vocab combination	.492
Vocab word formation	.569
Meaning	.639
Semantic relations	.508
Listening ability	.945
Listening intensively for detail	.633
Listening inferencing	.844
Listening to identify main idea	.926

Factor analyses were also conducted to explore the relationship between listening subskills and the macro skills, but in every case only one factor emerged (as happened in the case of reading, see Chapter 9) accounting for substantial amounts of variance (57 per cent to 77 per cent). Interestingly, even when vocabulary subskills were added to the analysis, no second factor emerged (Table 10.16 shows the factor structure), suggesting a degree of relationship between vocabulary and listening subskills.

Summary and conclusion

In this chapter we have discussed the construct of listening, described the DIALANG Listening test specifications, exemplified Listening items, and analysed their performance, as well as the performance of pilot test-takers.

We found that learners' listening ability seems to improve with the length of time they have been learning English, and the frequency with which they use the language. It also increases by level of education, but not with age. The ability to understand spoken English also varied by mother tongue, with speakers of languages closely related to English performing better than speakers of less related languages. There was no significant difference between males and females in listening ability, but a learner's self-assessed listening ability did indeed correspond to their listening ability as measured by the DIALANG tests.

We saw that one's abilities in listening subskills did not relate to a learner's CEFR level or to their performance on the Listening test as a whole. Even low-level learners are able to answer some questions that test inferencing abilities, as well as items testing the ability to understand main ideas.

There are inevitably many questions that remain to be answered, partly because the design of the piloting of DIALANG test items did not allow one to explore the relationship among the full range of skills

tested by DIALANG. Such a project, which would inevitably be rather extensive, would, however, be very valuable in helping us both to understand better the relationship among the skills and subskills tested in DIALANG, and to explore how the Council of Europe levels can be related to each other across skills and partial competences.

It is clear that more research is needed into how the different language skills and aspects of knowledge develop. This is particularly true for listening, which is a relatively unresearched area. Notable is the fact that to date there has been little research into which subskills or abilities might enable more useful and accurate diagnoses of listening ability. Whilst the three subskills tested to date in DIALANG clearly have their uses, they do not appear to correspond to developments in listening ability, and so future research would do well to explore what variables better predict the development of such an ability. It is quite possible, as we shall discuss more generally in the final chapter, that tests of aspects of linguistic ability and knowledge directly related to understanding foreign language sounds – phoneme discrimination, understanding of stress and intonation, the ability to cope with a range of accents, both native and non-native, and the like – will be avenues worthy of further exploration. Buck (2001) will be a good source of inspiration for ideas of relevant variables to incorporate into the research.

Chapter 11: Writing

As in the previous two chapters, in this chapter we will first discuss the construct – in this case, writing in a foreign language – and then describe the construct that DIALANG attempts to measure, by examining the specifications for the test. We will then look at sample test items from Version 1 of DIALANG, before presenting the results of the piloting. Next we explore how learners who completed the English Writing tests performed and the relationship between their performance and various background variables. Finally we will briefly discuss what more needs to be done before we can be confident that we are able accurately to diagnose a learner's developing writing ability in a foreign language.

The construct

The ability to write in a foreign language is one of the most difficult to develop, and one of the least often taught, at least in secondary schools. Nevertheless, and especially at tertiary level, particularly in universities, students whose first language is not that of the host institution usually need to be able to write in the language that is the medium of instruction, since much of the evidence for learning and achievement at university level is provided in written form. Because of this, a considerable amount of research has been devoted to developing an understanding of what is involved in writing, especially in academic contexts, in both one's first and a second or foreign language.

Weigle (2002) gives a very comprehensive account of the research and theory in this area. She follows much tradition in describing the different text types that are typically required of writers, she explores the differences between writing and speaking, between writing in one's first language and in a second or foreign language, and she pays considerable attention to writing as a social and cultural phenomenon, as well as to cognitive aspects of the writing process. She considers a whole array of variables that are thought, or that have been shown, to affect the assessment of writing, including task variables like content, audience and purpose, as well as more traditional variables like the rater and the rating

scales or other procedures used to evaluate the writing products, and also context variables and variables within the test taker him- or herself. Test developers or teachers wishing to understand better how to construct both tasks and scales for writing assessment are strongly recommended to read her Chapters 5 and 6 in particular.

Most writing assessments and research into, and textbooks about, writing assessment, are concerned with performance assessment. That is, they concentrate on how to get learners to actually produce relevant pieces of writing (the task variable) and how human raters can most validly and reliably evaluate the results (the rating criteria variable). In other words, when writers are assessed, they are usually asked to write something which is then evaluated by a trained human being,

Computer-based writing assesssment, of which DIALANG is but one example, is rather different from that, and will remain so for some time, until technology has advanced. When writing is assessed by computer, rather than by humans, the most that can happen at present is that the computer uses algorithms for automated text analysis to classify the writer's product in terms of (typically linguistic or content) variables that are known to correlate with human judgements of writing. The computer does not and cannot replace a human rater in this process, although it can produce analyses that will predict with reasonable accuracy the sort of judgements that human raters make. Nevertheless, such computer scoring requires sophisticated programming, and models for algorithms relevant for the assessment of writers of foreign languages, rather than native writers, are still crude in the extreme.

An alternative approach to this surrogate rating procedure has been in use for several decades, and although it has in some contexts been replaced by performance assessment of the sort described above, it is still used for many purposes, and that is the indirect testing of writing.

For indirect tests, test constructors develop tasks which attempt to assess the sorts of things that writers have to do when writing, for example use vocabulary and syntax accurately, organize texts intelligibly, use appropriate registers, spell and punctuate accurately and so on. These different components of writing ability, as they are often called, are assessed indirectly through traditional objectively assessable techniques like multiple-choice.

The justification for the use of such indirect techniques was traditionally that the results correlated well with direct measures of writing ability rated by human beings – such, for example, was for decades the justification for the indirect tests of writing ability used by Educational Testing Service in its TOEFL test until the mid-1980s. Such justifications may be somewhat dubious in these days of performance testing, especially in the case of high-stakes proficiency testing, where it is regarded as important to simulate as closely as possible the circumstances in which a writer will be writing for real. Yet in the case of diagnostic tests, which seek to identify relevant components of writing

ability, and assess writers' strengths and weaknesses in terms of such components, the justification for indirect tests of writing is more compelling. Indeed this is a theme to which we will return in the final chapter of this volume.

In the case of DIALANG, which is both diagnostic and computer-delivered as well as computer-scored, the need for direct tests of writing ability is less self-evident in current circumstances, with current technology, and indirect tests of writing ability are justified theoretically as well as practically.

The design of DIALANG Writing tests

The DIALANG Assessment Specifications (DAS) describe in some detail the construct that the Writing tests are intended to measure. These specifications are in part derived from the CEFR, but the CEFR is much more relevant to performance tests of writing than to the diagnosis of writing ability. DIALANG therefore sought other sources for insight into what sort of components might be involved in writing that could provide useful diagnostic information.

Given that there is not much empirical evidence regarding how the ability to write in a foreign language develops, much less how to diagnose its development and progress, DIALANG recognized the impossibility of developing specifications for each CEFR level separately. Instead, the DAS merely indicates to item writers the sorts of things they might take into consideration when writing test items aimed at the various CEFR levels.

The Introduction to the DAS for Writing states:

> *The test items cover three aspects of writing: register/appropriacy, accuracy, and textual organisation. The task types used are multiple-choice questions, short-answer questions and two kinds of gap-filling tasks.*

The DIALANG assessment system focuses on communicative writing for:

- giving and exchanging information,
- persuading/arguing a point of view, and
- using writing for social interaction.

Table 11.1 presents the domain specification used in DIALANG. The table was developed originally by Anneli Vähäpassi for the IEA International Study of Writing (Gorman *et al.*, 1988). It presents a model that usefully combines aspects of cognitive processing, the purpose of writing, the content of writing and the audience of writing.

The cells of the table contain various text types. The classification is, however, not simple and clear-cut, since one text may share elements of

Table 11.1 Generic domain specification for writing. For more information see Gorman *et al.* 1988.

Cognitive Processing / Primary content	Primary Audience	Reproduce		Organize/Reorganize		Invent/Generate	
Dominant intention/ Purpose		Facts	Ideas	Events	Visual images, facts, mental states, ideas	Ideas, mental states, alternative worlds	
To learn (metalingual)	Self	copying, taking dictation		retell a story (heard or read)	note, resume, summary, outline, paraphrasing	comments in book margins metaphors, analogies	
To convey emotions, feelings (emotive)	Self, others	stream of consciousness		personal story personal diary personal letter	portrayal	reflective writing, personal essays	The traditional literary genres and modes can be placed under one or more of these five purposes
To inform (referential)	Others	quote	fill in a form	narrative report, news, instruction, telegram, announcement, circular	directions, description, technical description, biography, science report/experiment	expository writing, definition, academic essay/article, book review, commentary	
To convince/ persuade (conative)	Others	citation from authority/expert		CV, letter of application	advertisement, letter of advice, statement of personal views, opinions	argumentative/ persuasive writing, editorial, critical essays/article	
To entertain, delight, please (poetic)	Others	quote poetry and prose		given an ending create a story, create an ending, retell a story	word portrait or sketch	entertainment writing, parody, rhymes	
To keep in touch (phatic)	Others	greeting card		postcard	personal letter	humorous greeting	

different types. Moreover, this typology does not necessarily reflect the 'difficulty' of the various text types *per se*; rather, it is used as an aid to cover adequately the domain of writing.

Although there are different ways of classifying text types, the classification provided by Werlich (1976, 1988) is the main source for the DIALANG approach. The main categories (descriptive, narrative, expository, argumentative and instructive) presented in Chapter 9 on reading were also taken into account when creating Writing items.

The purposes of writing employed in DIALANG are adopted from the Council of Europe Framework, where they appear under 'functional competence'. Those appropriate for writing are:

- giving and seeking factual information;
- expressing and finding out attitudes and opinions;
- persuading;
- socializing.

Another aspect of the domain of writing has to do with the **roles** (in a broad sense) of the writer and the intended audience. Roles/role relationships are affected by a number of factors including:

- difference in social status;
- difference on the novice–expert status dimension;
- personal vs public context;
- situationally assigned roles such as:
  - host/hostess vs guest(s);
  - socializing roles: invitations, written replies, thank-you letters;
  - citizen vs authority; client vs provider of public services;
  - meeting roles: chairperson, secretary, meeting participant, observer, etc.;
  - reporting roles: minutes, notes, rapporteur's reports, etc.

See also the Council of Europe Framework (2001).

In DIALANG, indirect writing items assess users' skills in recognizing elements of appropriacy, accuracy and coherence in writing. These items focus on the users' *potential* to write, rather than on the actual ability to produce written texts. The writing tasks in the current DIALANG system tap users' sensitivity to issues that might create problems in writing due to lack of target language awareness. This form of assessment is used because of its ability to produce reliable estimates in skills related to writing. The collection of Experimental Items developed by DIALANG (see Chapter 15) includes one potential task type for assessing productive writing, but this item type requires comprehensive trialling (see Chapter 16 for an account of research into the value of such an Experimental Item). Indirect writing assessment may focus on mechanics, such as accuracy (e.g. spelling, abbreviations, grammar), setting the tone of

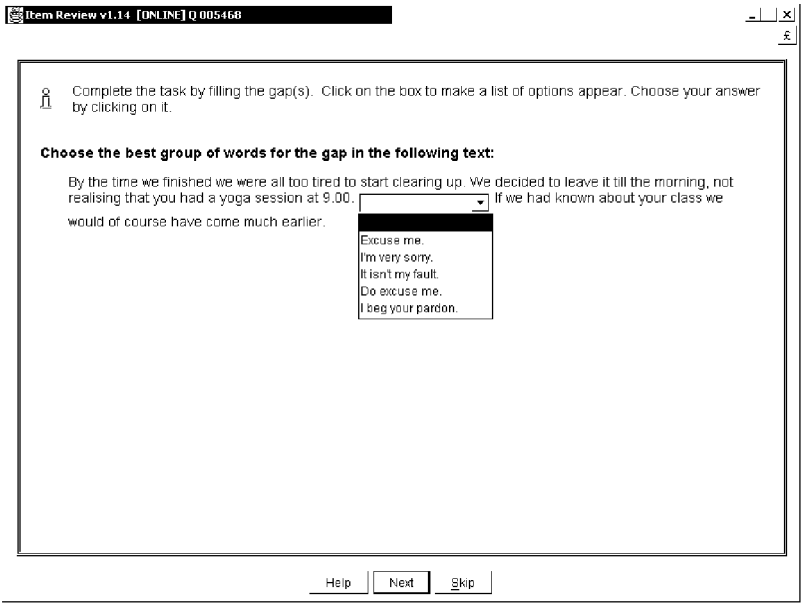

Figure 11.1 Register/appropriacy

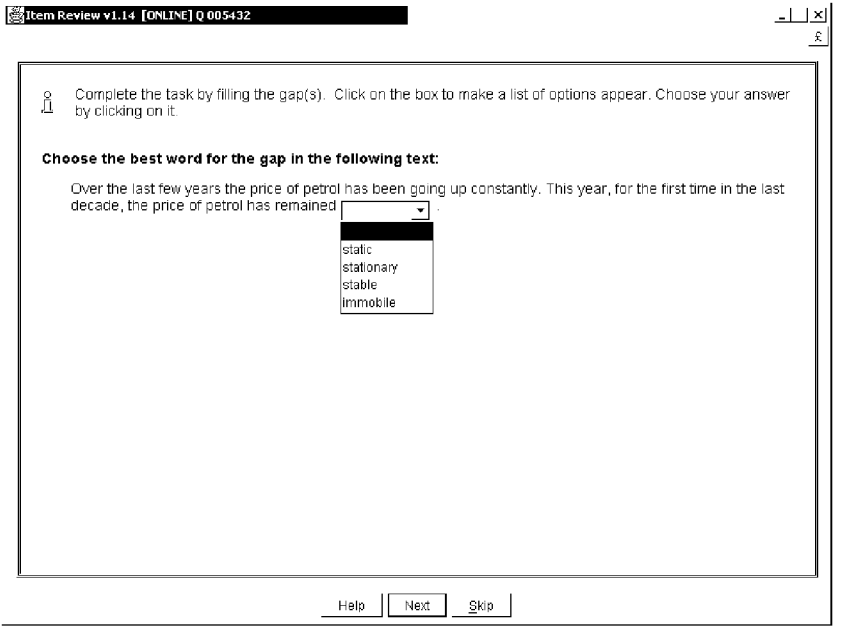

Figure 11.2 Accuracy (vocabulary)

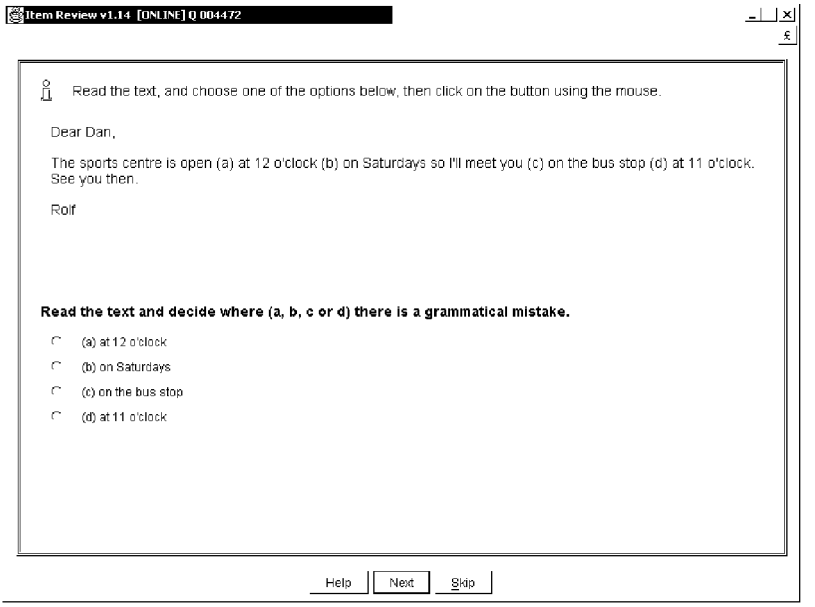

Figure 11.3 Accuracy (syntax)

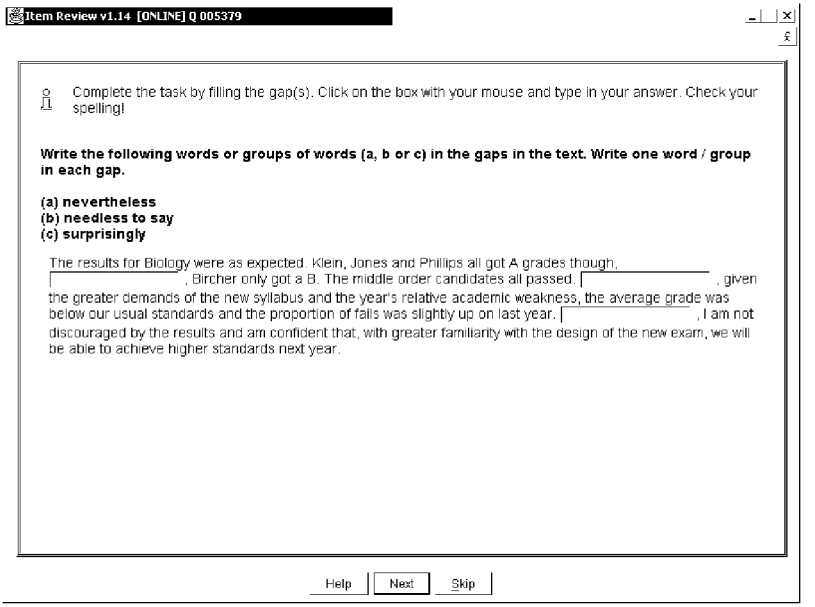

Figure 11.4 Textual organization

the text (e.g. appropriacy and convention in letter writing, differences between formal and informal register), or text construction skills, such as the ability to detect coherence and cohesion markers.

Test items

In this section, in Figures 11.1–11.4 I present a range of different items, as realized in the English Writing test in DIALANG, Version 1.

The results of piloting

On the whole, piloting of English Writing items was very successful: 54 out of 60 items survived into Version 1. All six rejected items were testing 'knowledge of accuracy' (out of 28 piloted). In these cases, there were doubts about what was being tested, and in two cases there were too many acceptable answers.

Table 11.2 shows the number of surviving items by item type, and tests to see whether there are significant differences in difficulty across item types. It is noticeable that there are many more text-entry items in the Writing tests than in other tests, and the drop-down type of multiple-choice is preferred to the more traditional type. A one-way ANOVA showed no significant difference.

Table 11.3 shows how many items survived piloting, classified according to their author's best guess (ABG) as to their level, and contrasts ABG with actual empirical difficulty as established by the piloting.

Although a one-way ANOVA showed a significant difference in mean scores across the ABG levels, *post-hoc* contrasts show a significant difference ($p = .017$) between A and C levels only. If we crosstab the

Table 11.2 Difficulty by item type

Item type	N	Mean	Min	Max	Std Deviation	
Drop-down	22	−.097	−.98	1.03	.459	
Multiple-choice	6	−.256	−.64	.21	.338	
Text-entry	26	.141	−.55	1.31	.442	F = 2.867 P = .066

Table 11.3 Difficulty by Author's Best Guess

Item level	N	Mean	Min	Max	Std Deviation	
A	17	−.256	−.867	.34	.324	
B	22	.065	−.98	1.31	.466	
C	15	.194	−.44	1.03	.462	F = 4.887 P = .011

Table 11.4 Difficulty by subskill

Subskill	N	Mean	Min	Max	Std Deviation	
Knowledge of accuracy (grammar/ vocab/spelling)	22	.192	−.639	1.307	.499	
Knowledge of register/ appropriacy	15	−.141	−.983	.888	.447	
Knowledge of textual organization	17	−.124	−.641	.477	.316	F = 3.629 P = .034

Table 11.5 Subskill by CEFR level

CEFR	A1	A2	B1	B2	C1	Total
Accuracy		2	11	7	2	22
Register/appropriacy	2	2	10	1		15
Textual organization		5	10	2		17
Total	2	9	31	10	2	54

empirically established CEFR level by ABG, we find no significant association ($\chi^2 = 24.877$, p = .206).

If we then examine the items by subskill to see whether items testing one subskill are more difficult than those testing other subskills, Table 11.4 shows the results.

Although a one-way ANOVA just reaches significance, *post-hoc* contrasts show no significantly different pairs. Moreover, if we crosstab subskills by empirically established CEFR level (Table 11.5), we find no significant association ($\chi^2 = 14.722$, p = .065). Thus the evidence is at best weak that there is a relationship between subskill and the level of the items.

Writing ability

In earlier sections we examined the Writing test items, and their performance during piloting. In this section, we will examine the scores of those learners who took English Writing tests as part of the piloting. Readers will recall that pilot test-takers were assigned at random to one of four pilot Writing booklets. Two of these booklets also contained Grammar items, and two contained Vocabulary items. This means that

Table 11.6 Descriptive statistics for writing ability

N	Minimum	Maximum	Mean	SE Mean	Standard Deviation
735	−1.227	2.094	.198	.018	.483

Table 11.7 Writing ability by mother tongue

Mother tongue	Mean	N
Danish	.540	42
Icelandic	.513	33
Swedish	.394	14
Norwegian	.394	36
German	.350	151
Dutch	.255	104
Italian	.248	10
French	.053	60
Portuguese	.053	11
Spanish	.041	41
Finnish	.029	121
Greek	−.025	5
Other L1s	−.044	107

we can compare learners' performance on Writing tests with their performance on Grammar and Vocabulary tests.

First I present the descriptive statistics for performances on the Writing Test (Table 11.6). I then analyse results according to background variables. Finally I look at the relationships among learners' skills and subskills.

Writing ability by mother tongue

A one-way ANOVA showed significant overall differences ($F = 10.995$, $p = .000$) with Danish, Icelandic, German, Norwegian and Swedish showing significant advantage. Data are presented in Table 11.7.

Writing ability by sex

Females perform significantly ($p = .000$) better than males (Table 11.8).

Table 11.8 Writing ability by sex

Sex	Mean	N
Female	.257	446
Male	.107	289

Table 11.9 Writing ability by age

Age	Mean	N
Under 18	−.063	23
18–25	.165	462
26–35	.242	125
36–45	.319	71
46–55	.351	44
56–65	.448	7
Over 65	−.331	3

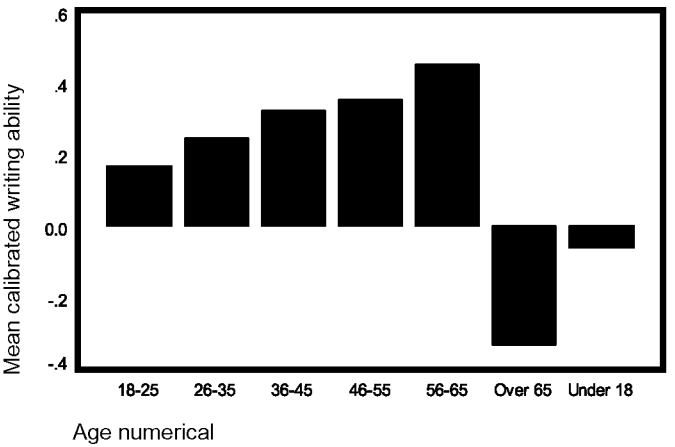

Figure 11.5 Writing ability by age

Writing ability by age

See Table 11.9. A one-way ANOVA shows significant overall differences ($F = 4.149$, $p = .000$), with the under 18s being significantly weakest at writing. Figure 11.5 shows the results graphically, but the size of the samples should be borne in mind.

Writing ability by educational level

A one-way ANOVA showed significant differences overall in writing ability across educational level ($F = 20.558$, $p = .000$), with university being unsurprisingly higher than secondary or primary (Table 11.10).

Writing ability by length of time learning English

Results are shown in Table 11.11. A one-way ANOVA showed significant differences overall ($F = 25.349$, $p = .000$), with more than 10 years' experience scoring highest. Figure 11.6 shows the results graphically.

Table 11.10 Writing ability by educational level

Education	Mean	N
Higher (university)	.325	364
Other	.190	15
Secondary (general)	.170	159
Higher (non-university)	.168	96
Primary	−.164	32
Secondary (vocational)	−.199	69

Table 11.11 Writing ability by length of time learning English

Length of time	Mean	N
Less than a year	−.173	27
1–2 years	−.257	38
3–4 years	−.024	79
5–6 years	.127	140
7–8 years	.235	147
9–10 years	.232	157
More than 10 years	.495	147

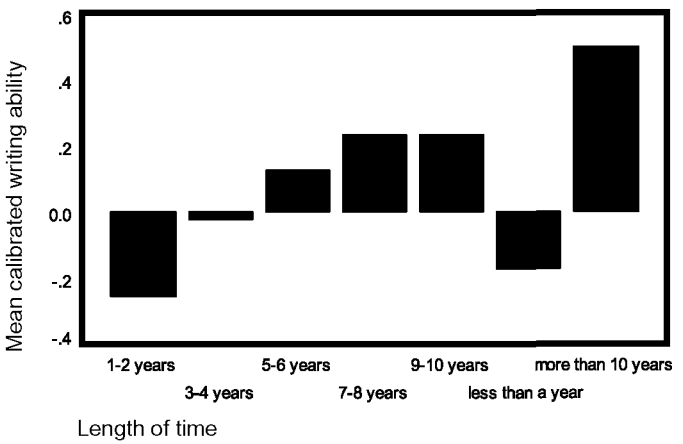

Figure 11.6 Writing ability by length of time learning English

Writing ability by frequency of use of English

Table 11.12 shows the results. A one-way ANOVA showed significant differences overall ($F = 17.808$, $p = .000$), but only everyday use had a significant advantage. Figure 11.7 presents the results graphically.

Table 11.12 Writing ability by frequency of use of English

Frequency of use	Mean	N
Cannot say	.101	71
Once a month or less often	.036	85
Once every two weeks	.046	67
Once or twice a week	.156	321
(Almost) every day	.429	191

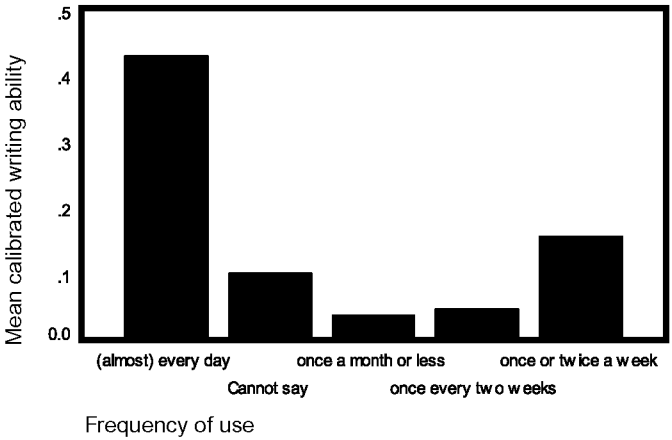

Figure 11.7 Writing ability by frequency of use of English

Writing ability by overall self-assessed writing ability

See Table 11.13. A one-way ANOVA showed significant differences overall ($F = 74.939$, $p = .000$), with self-assessed higher CEFR levels corresponding to higher scores for writing. These results provide support for the argument that indirect tests of writing possess a degree of validity. Figure 11.8 shows the results graphically.

Table 11.13 Writing ability by overall self-assessed writing ability

SA level	Mean	N
A1	−.250	60
A2	−.237	74
B1	.055	221
B2	.317	199
C1	.548	125
C2	.607	56

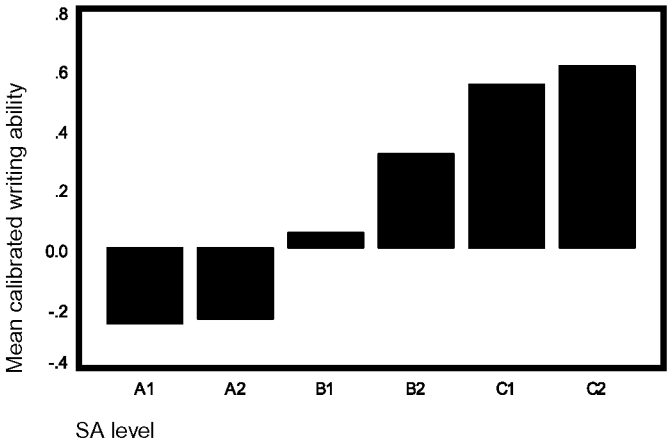

Figure 11.8 Writing ability by self-assessed writing ability

Relationship between skills and subskills

Test-takers assigned to a Writing test were also given either a test of Grammar or a test of Vocabulary. We can therefore compare performance on these tests, both at macro skill and subskill levels. The correlation between Writing and Grammar was .77, sample size n = 366, and between Writing and Vocabulary .79, sample size n = 369. These correlations are substantial and show that writing as measured in DIALANG is closely related to grammatical and lexical knowledge.

Table 11.14 shows the correlations between writing subskills and the macro Writing skill, which is substantial and undifferentiated.

Table 11.14 Writing subskills and macro Writing skill

	Writing ability
Accuracy	.875
Register and appropriacy	.820
Textual organization – coherence/cohesion	.813

Table 11.15 Writing subskills and Grammar (n = 366)

	Grammar ability
Accuracy	.744
Register and appropriacy	.580
Textual organization – coherence/cohesion	.618

Table 11.16 Writing subskills and Vocabulary (n = 369)

	Vocabulary ability
Accuracy	.780
Register and appropriacy	.692
Textual organization – coherence/cohesion	.556

Table 11.17 Writing subskills and grammatical ability

	Factor 1
Accuracy	.873
Register and appropriacy	.717
Textual organization – coherence/cohesion	.721
Grammar ability	.845

Table 11.18 Writing skills and vocabulary ability

	Factor 1
Accuracy	.881
Register and appropriacy	.790
Textual organization – coherence/cohesion	.644
Vocabulary ability	.880

Table 11.15 shows the relationship between writing subskills and grammar knowledge and Table 11.16 shows the relationship between writing subskills and vocabulary knowledge.

Clearly writing accuracy and grammar and vocabulary are closely related, more than register/appropriacy and coherence/cohesion are related to grammatical knowledge or vocabulary knowledge.

Factor analyses were conducted of the relationship between writing subskills and the macro skills of Writing, Grammar and Vocabulary (see Table 11.17 and Table 11.18). However, in every case, only one factor emerged, accounting for between 72 per cent and 77 per cent of the variance. Interestingly, the three writing subskills showed high loadings on a factor including grammar ability, and also on a factor including vocabulary ability.

Clearly the writing subskills tested in DIALANG are related to both grammatical and lexical abilities, which suggests that diagnostic measures might be valuable that take account of all three skills (Writing, Vocabulary and Grammar) and include subskills for grammar and vocabulary as well as for writing.

Summary and conclusion

After a brief discussion of the construct of writing ability, and a justification for the inclusion in DIALANG of indirect tests of writing ability, we examined the DIALANG specifications for tests of writing and presented a number of indirect writing tasks. The performance of the tasks was examined, to show that there was no significant association between the CEFR level of an item and the subskill it was measuring.

Learners' scores on the English Writing test were compared with background variables, to show that mother tongue, sex, age, educational level, length of time learning English and frequency of use of English all had a significant effect on a learner's writing ability as measured by DIALANG. Happily, learners' self-assessed writing ability showed a significant relationship with their empirically calibrated CEFR level, even though the tests were indirect in nature. This gives some confidence as to the value of indirect Writing items as diagnostic measures of the ability to write in a foreign language.

An analysis of the performance of the Writing items in terms of relationships among skills and subskills showed a substantial amount of variance in common across items testing Writing, Grammar and Vocabulary. Whilst in a sense this is not surprising, given that the Writing items were indirect in nature, and therefore arguably more grammatical and lexical in content than more direct measures of writing ability might have been, it also suggests that there might well be considerable value in future research exploring the contribution to writing ability that is played by more traditionally linguistic measures of knowledge and accuracy than is often currently thought to be the case. Given that the subskill-based diagnostic items showed little relation to CEFR levels, there might be an argument for exploring the extent to which writing ability can be shown to develop in parallel with development in a learner's grammatical and lexical abilities, as well as in their ability to organize their ideas in text, and express themselves appropriately in writing.

Chapter 12: Grammar

In this chapter, we first discuss what is meant by the construct of 'grammar', then I describe DIALANG's approach to the assessment of grammar (also known in DIALANG as 'Structures') and exemplify with a number of items taken from Version 1 of DIALANG. We next analyse the performance of the items that were piloted, and then look at how learners taking Grammar tests performed and how their scores related to background variables. Finally we examine the relationship between grammatical subskills and macro skills of Grammar, Reading and Writing, and draw tentative conclusions for further research and development.

The construct

A knowledge of the grammar of a language has always been regarded as central to the development of foreign language proficiency. What has varied over time is exactly what is meant by 'knowledge of grammar'. Traditionally, the ability to translate texts from and into the foreign language using target syntax and morphology accurately was regarded as an indicator of grammatical knowledge. In addition, the ability to recite or recall the rules for the construction of grammatically accurate forms and sentences was also believed to be a crucial component of grammatical ability.

More recently, less emphasis has been placed on the ability to conjugate verbs, decline nouns and transform phrases and clauses into grammatically appropriate forms in the given context. Associated with a de-emphasis on the importance of form, and an increased stress on the importance of meaning, together with an insistence that appropriacy of use is at least as important as accuracy of usage, communicative language teaching and testing has tended to emphasize the important relationship between grammatical forms and their meaning. This has had the implication that assessment has tried to devise ways in which a learner is tested on his or her ability to express the meaning required by the task or the context in grammatically appropriate ways.

Purpura (2004) provides an excellent overview of the ways in which the grammatical construct has been viewed over time, and describes at some length, with numerous examples, how the ability to use grammatically appropriate and accurate forms might best be conceptualized and assessed.

Unlike the language use skills we have examined in previous chapters, however, the explict assessment of grammar has never really succumbed to the so-called communicative revolution, at least when it has been assessed at all. Certainly, many writers have proposed that there is no need to focus explicitly on grammar, since grammar is at the heart of language use, and therefore testing a student's ability to read, to listen, to write and to speak in a foreign language is necessarily also testing his or her ability to use the language grammatically. Thus, if learners cannot understand the syntax of a written text, they will not be able to understand its meaning and the same, *mutatis mutandis*, with spoken texts. In the case of the assessment of writing and speaking, performance testing implies that raters must pay attention to key components of the ability being evaluated, and it has been common practice for rating scales to include an explicit assessment of grammar as part of the assessment of performance. Nevertheless, in many parts of the world, and in many testing systems, some test of grammar has been retained alongside tests of the four skills. Such tests of grammar often go under the name of Use of Language, rather than Grammar, yet the constructs on which they are based are indistinguishable from tests of the ability to manipulate grammatical forms in relation to meaning.

Interestingly also, when attempts have been made to diagnose foreign language ability, these have tended to focus upon grammar. The diagnostic grammar tests developed by Tim Johns at Birmingham University in the UK in the 1970s and early 1980s are a case in point: Johns felt that there was considerable value in assessing, for diagnostic purposes, whether a student could manipulate grammatical forms accurately and use them appropriately in suitable contexts (personal communication, 2003).

The assessment of grammatical abilities in DIALANG is arguably not far removed from much normal practice in foreign language testing more generally, at least insofar as there has been a focus on the use of the language. And indeed, since explicit grammar tests did not normally require long texts, but could quite often be adequately contextualized using short texts with little contextual constraint (especially bearing in mind the often-expressed need to separate the testing of grammar from the assessment of reading), then even computer-based tests of grammar have not been markedly different in appearance from paper-and-pencil based tests of the same construct.

Thus the DIALANG tests of grammar, which are largely sentence-based, are perhaps less conservative than might be thought to be the case for other skills. What is different is the explicit attempt to identify

aspects of grammar that are relevant diagnostically, and to report these subskills separately.

The DIALANG Assessment Specifications (DAS) describe in some detail the construct that the Grammar tests are intended to measure. These specifications are only in part derived from the CEFR, because the CEFR does not contain any description of a grammar construct or an underlying theory. This is unfortunate, since the diagnosis of grammatical ability must involve some notion of what it means to know and use the grammar of a language. Be that as it may, the CEFR is neutral as to which language is to be taught or tested, and therefore has little to say about the nature of grammar in particular languages. DIALANG therefore had to turn to other sources for insight into what sort of useful diagnostic information could be derived from assessment.

Despite the recent upsurge of interest in research into second language acquisition, there is still not much empirical evidence regarding how foreign language learners develop their grammatical competence in languages other than English (and even in English the evidence is very incomplete), and so DIALANG had to recognize the impossibility of developing specifications for each CEFR-related level separately.

However, what DIALANG did do, as part of the development of CEFR-related standard-setting procedures for grammar, was to begin to develop a set of self-assessment statements (as described in Chapter 6) which might eventually be used, both to inform test specifications and to contribute to the development of self-assessment of grammatical ability. Therefore, towards the end of this chapter, we will examine the empirical results of the standard-setting procedures, to see what light they might be able to throw on the development of grammatical ability.

The construct of grammar in DIALANG

There are a great number of grammatical models, ranging from traditional to functional, from tagmemic to generative-transformational, from systemic to lexical-functional, and so on. The role of grammar and categories of grammar may vary considerably depending on the approach chosen. Moreover, it is often difficult to draw a clear distinction between grammar and vocabulary, and languages contain numerous features that occupy an area between lexis and grammar. On the morphological level, derivation and compounding may be seen to lie between grammar and vocabulary, though they will usually be considered to belong to vocabulary, whereas inflection may be considered a natural component of grammar. Test Development Teams followed the conventions of the teaching traditions in their language in deciding which section these components should belong to.

In the Grammar tests the items measure the user's ability:

1 to understand and use morphology;
2 to understand and use syntax.

The experience of designing indirect writing items (see Chapter 11) showed that coherence and cohesion items were of a slightly different order than the other grammatical categories, and closer to those aspects of language assessed in the DIALANG Writing section. Therefore, the decision was made to assess text grammar in the Writing section, and to concentrate the grammar/structure items on morphological and syntactic features of each language. The DIALANG test of grammar is thus mainly focused on the traditional categories of morphology and syntax. This does not mean that functions and notions have not been taken into account at all. However, the main criterion for inclusion in the Grammar test has been adequate coverage of the traditional grammatical categories.

However, the project recognized that it was not possible to create a detailed framework of structures which would be appropriate for all 14 languages. Therefore, the Test Development Teams adapted the list below for the needs of their own language. The adaptations involved specifying further the focus of some categories, and adding new categories to cover specific features of individual languages.

Outline of grammatical categories

1 MORPHOLOGY
1.1 NOUNS
1.1.1 inflection – cases
1.1.2 definite/indefinite – articles
1.1.3 proper/common
1.2 ADJECTIVES AND ADVERBS
1.2.1 inflection
1.2.2 comparison
1.3 PRONOUNS
1.3.1 inflection
1.3.2 context
1.4 VERBS
1.4.1 inflection: tense, mood, person
1.4.2 active/passive voice
1.5 NUMERALS
1.5.1 inflection
1.5.2 context
1.6 OTHER
2 SYNTAX
2.1 ORGANIZATION/REALIZATION OF PARTS OF SPEECH
2.1.1 word order: statements, questions, exclamations
2.1.2 agreement
2.2 SIMPLE vs COMPLEX CLAUSES
2.2.1 coordination

2.2.2 subordination
2.2.3 deixis
2.3 PUNCTUATION
2.4 OTHER

There are 19 categories mentioned above, 17 named ones and two 'others', one for morphology and one for syntax. For each language the 10–15 most relevant categories were selected. In some cases, additional categories were found to be necessary.

As far as possible, the use of grammatical metalanguage was avoided in instructions to users, so that items do not test metalinguistic knowledge.

To sum up, the items in the Grammar section:

1 cover at least the main types of grammatical structures within each category (e.g. all important types of verb);
2 range from very basic and easy/common structures or uses of structures to difficult structures and/or rarer uses of these structures;
3 allow users to demonstrate their ability both to understand and to produce the relevant structures;
4 are given through a variety of different task types.

Language-specific modifications and comments

The Test Development Teams largely followed the generic framework proposed above. The categorization is based on traditional grammars, with a main division between morphological and syntactic items. This compromise solution for the specifications was chosen because sufficient research did not exist to suggest another categorization applicable to all the 14 languages of DIALANG. After this main decision was made, the Test Development Teams tried to follow the general specifications, but amended the system proposed in three main ways:

- they omitted categories which did not concern their language or were thought trivial;
- they added categories or sub-categories which were important for their language;
- they modified the focus of some categories, again to suit the grammar of their language.

The teams added their own categories especially for areas which cover both morphology and syntax. The most frequent additions and amendments concerned adverbs and adverbials: their placement and use in simple and complex sentences seems to be a salient learning point in most European languages. Another frequent addition was the category of phrasal verbs. The teams also considered that prepositions and postpositions deserved a category of their own.

Several teams reported that they had created sub-categories within each DIALANG category. They did this to help the item writers' work, and provide sets of items which would be potentially useful for giving detailed feedback to learners. The categories stemmed from popular coursebooks and teachers' experience with learners.

Test items

In this section, we exemplify the various task types used in DIALANG, taking sample items from the English Grammar test in Version 1 of DIALANG. The English team devised grammar items according to 12 different categories, within a more general division into Morphology and Syntax. We give only a few samples, shown in the screenshots in Figures 12.1–12.5.

Analysis of the results of piloting of DIALANG Grammar tests

As a result of piloting, 56 out of 60 Grammar items (93 per cent) survived into Version 1. Four items were rejected, one testing the 'comparison of adjectives and adverbs' (out of five piloted), one testing 'morphology' and two testing 'syntax – punctuation'. Table 12.1 (p. 178) shows the number of each item type in Version 1, and their mean empirical difficulty.

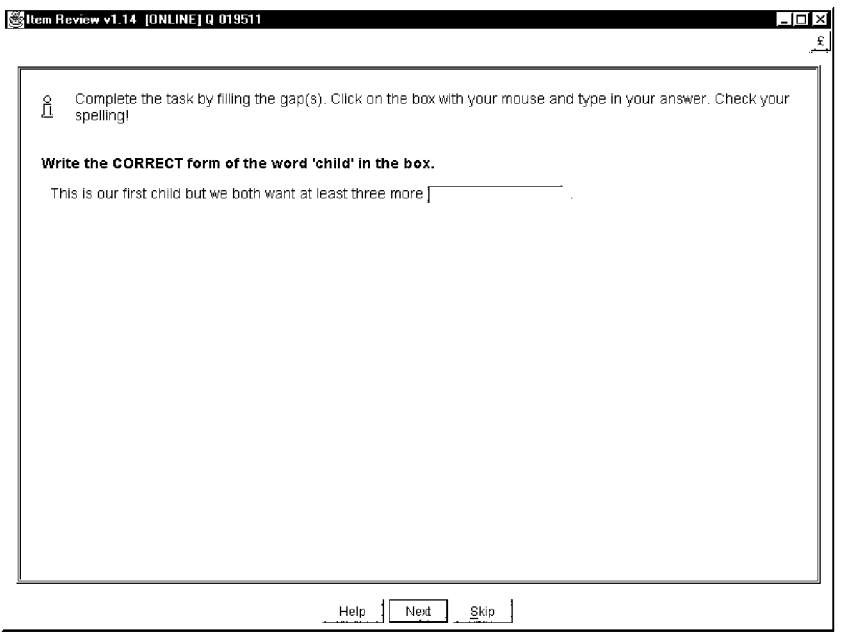

Figure 12.1 Proper and common nouns

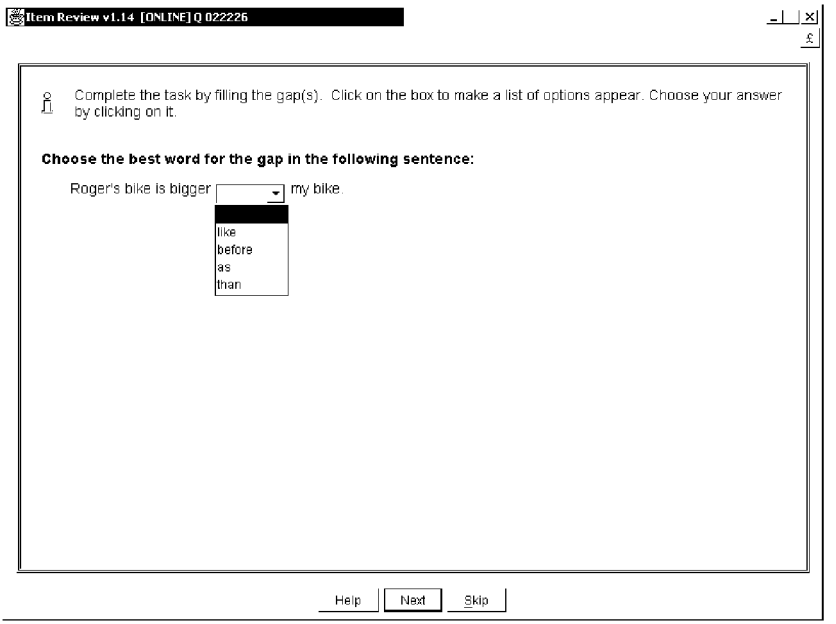

Figure 12.2 Comparison of adjective and adverb

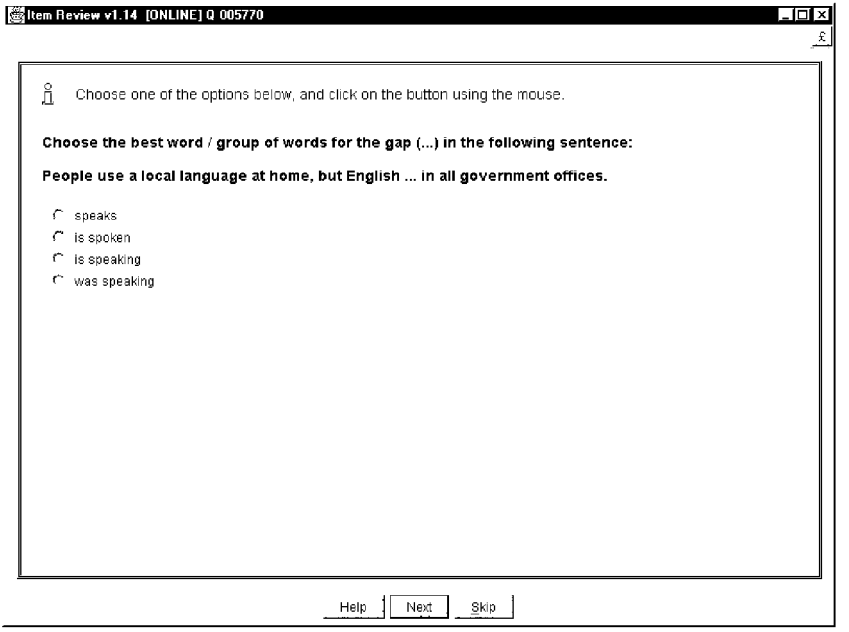

Figure 12.3 Verbs active and passive

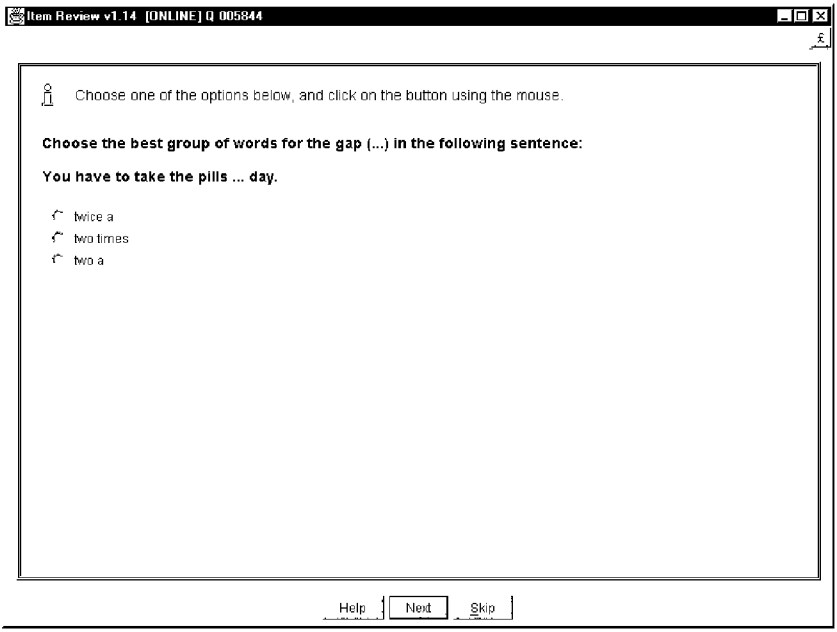

Figure 12.4 Morphology – numerals

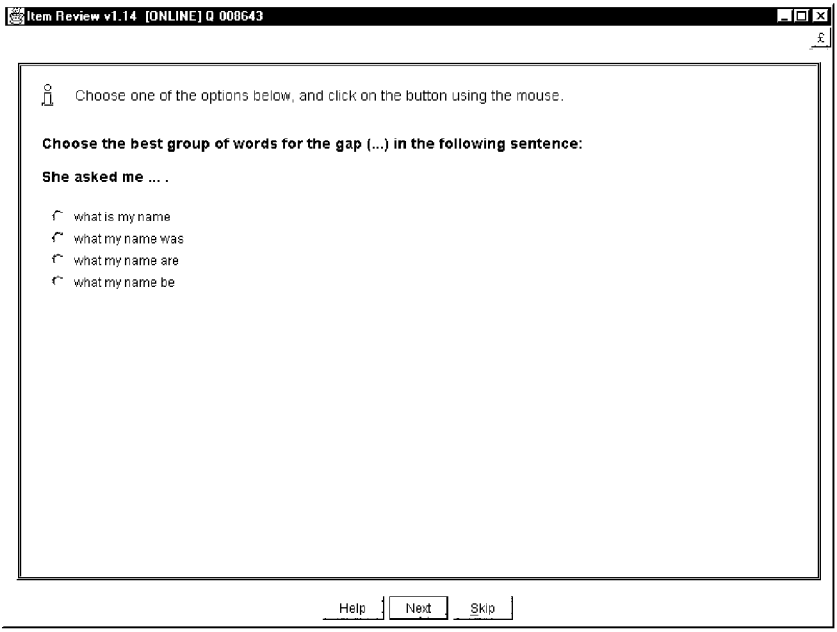

Figure 12.5 Syntax organization

Table 12.1 Difficulty by item type

Item type	N	Mean	Min	Max	Std Deviation	
Drop-down	9	−.259	−.839	.522	.4011	
Multiple-choice	32	.024	−.835	2.011	.5726	
Text-entry	15	.311	−.500	1.070	.5312	F = 3.274 P = .046

Table 12.2 Difficulty by Author's Best Guess levels

Item level	N	Mean	Min	Max	Std Deviation		
A	17	−.058	−.606	1.047	.4806		
B	23	.063	−.839	1.546	.5929		
C	16	.165	−.835	2.011	.6037	F = .642	P = .530

Although ANOVA just reaches significance, there are no significantly different pairs.

Table 12.2 shows a comparison between the Author's Best Guess at an item's level, and its empirical difficulty.

There are no significant differences across the 'levels'. In other words, authors failed to predict the difficulty of the items they wrote. When we crosstab CEFR level with ABG, we find the association is just significant ($\chi^2 = 32.406$, $p = .039$) but there are many cells (90 per cent) with an expected frequency below 5.

Table 12.3 reports the mean difficulties of the various subskills tested in the Grammar test. A one-way ANOVA reveals no significant difference in mean scores by subskill.

When we try to crosstab subskills by CEFR level (Table 12.4), we find, not surprisingly given the number of subskills and items, that there is no significant relationship ($\chi^2 = 41.311$, $p = .688$).

Grammatical ability

So far we have examined the Grammar test items, and their performance during piloting. In this section, we examine the scores of those learners who took English Grammar tests as part of the piloting. Those who took the pilot tests were assigned at random to a test booklet, four of which contained Reading and Grammar tests, and two of which contained Writing and Grammar tests. This means that we can compare learners' performance on Grammar tests with their performance on either Reading tests or on Writing tests.

Before I present an analysis of results on the Grammar test according to background variables and in relation to other skills, I present the

Table 12.3 Difficulty of subskills

Subskill	N	Mean	Min	Max	Std Deviation	
Morphology – Adjectives and Adverbs – inflection	1	−.427	−.43	−.43		
Morphology – Adjectives and Adverbs – comparison	4	−.197	−.54	.06	.2576	
Morphology – Nouns – definite/ indefinite articles	2	−.339	−.37	−.31	.0396	
Morphology – Nouns – proper/ common	4	.222	−.03	.47	.2743	
Morphology – Numerals – inflection	8	.066	−.84	1.55	.7778	
Morphology – Others	6	.265	−.53	2.01	.9155	
Morphology – Pronouns – context	6	.102	−.50	1.05	.5839	
Morphology – Verbs – active/passive voice	4	.209	−.22	1.04	.5775	
Morphology – Verbs – inflection tense, mood, person	6	.477	−.34	1.07	.4780	
Syntax – Organization/ realization of parts of speech – word order statements, questions	10	−.071	−.52	.30	.2821	
Syntax – Punctuation	4	−.252	−.84	.11	.4351	
Syntax – Simple sentences vs complex sentences coordination	1	−.606	−.61	−.61		F = .950 p = .504

descriptive statistics for scores on the Grammar test (Table 12.5). As in other chapters, scores were calibrated using IRT and are expressed in logit units, which range from a minus figure (for a relatively weak performance) to positive figures for stronger performances.

Table 12.4 Subskill by CEFR level

CEFR subskill	A1	A2	B1	B2	C1	Total
Comparison of adj and adv	1	3				4
Definite/indefinite articles		2				2
Inflections of adj and adv		1				1
Proper and common nouns		2	2			4
Morphology other	1	3	1		1	6
Pronouns – context	1	3	1	1		6
Verbs, tense, mood, etc.		1	4	1		6
Morphology – numerals	3	3	2			8
Verbs active and passive		3		1		4
Syntax punctuation	1	3				4
Simple and complex sentences	1					1
Syntax organization	1	6	3			10
Total	9	30	13	3	1	56

Table 12.5 Descriptive statistics for grammatical ability

N	Minimum	Maximum	Mean	SE Mean	Standard deviation
1084	−1.366	1.898	.739	.0152	.501

Table 12.6 Grammatical ability by mother tongue

Mother tongue	Mean	N
Norwegian	1.021	58
Icelandic	.950	31
Swedish	.937	21
Irish	.912	2
German	.905	225
Danish	.895	57
Dutch	.853	132
Italian	.831	13
Portuguese	.750	21
Finnish	.616	187
French	.571	91
Spanish	.542	81
Other L1s	.529	155
Greek	.388	10

Grammatical ability by mother tongue

Data are presented in Table 12.6. A one-way ANOVA showed significant overall differences ($F = 11.348$, $p = .000$) with paired contrasts often showing Finnish, French, Spanish and Other L1s significantly lower than other languages, and Norwegian, Icelandic, Swedish, Irish, German, Danish, Dutch, Italian and Portuguese forming a homogeneous, and higher, subset.

Table 12.7 Grammatical ability by sex

Sex	Mean	N
Female	.794	663
Male	.651	421

Table 12.8 Grammatical ability by age

Age	Mean	N
Under 18	.585	37
18–25	.749	682
26–35	.705	180
36–45	.748	108
46–55	.908	58
56–65	.825	17
Over 65	−.294	2

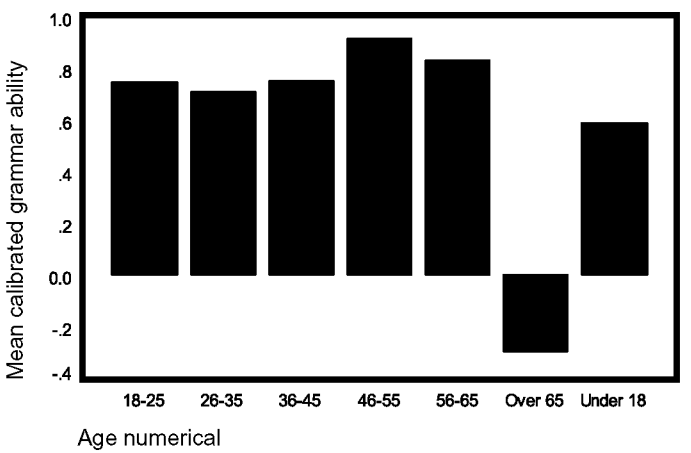

Figure 12.6 Grammatical ability by age

Grammatical ability by sex

See Table 12.7. Females outperformed males significantly ($p = .000$).

Grammatical ability by age

Table 12.8 shows the results. Age showed a significance difference in grammatical ability overall ($F = 3.367$, $p = .000$) but *post-hoc* tests revealed no significantly different pairs. Yet again, age seems not to be a major factor on its own. Figure 12.6 shows the results graphically.

Grammatical ability by educational level

See Table 12.9. Educational levels showed significant differences overall ($F = 17.450$, $p = .000$), with the main significant contrasts being between university and secondary/primary levels.

Grammatical ability by length of time learning English

Table 12.10 shows the results. A one-way ANOVA revealed significant differences overall ($F = 40.599$) and across most paired comparisons – Figure 12.7 presents the results graphically.

Table 12.9 Grammatical ability by educational level

Education	Mean	N
Higher (university)	.826	530
Secondary (general)	.754	232
Higher (non-university)	.744	144
Other	.694	25
Secondary (vocational)	.427	107
Primary	.383	46

Table 12.10 Grammatical ability by length of time learning English

Length of time	Mean	N
Less than a year	.460	36
1–2 years	.148	59
3–4 years	.462	118
5–6 years	.724	225
7–8 years	.771	219
9–10 years	.802	207
More than 10 years	1.0139	220

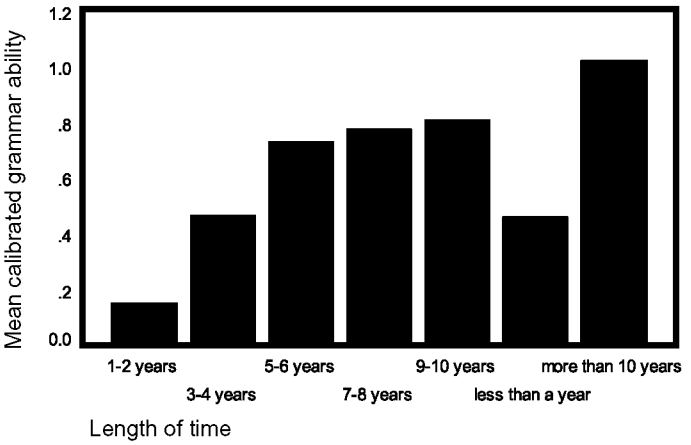

Figure 12.7 Grammatical ability by length of time learning English

Grammatical ability by frequency of use of English

Table 12.11 presents the data. A one-way ANOVA showed significant differences overall ($F = 15.914$, $p = .000$), with everyday use being always significantly higher than other contrasts. Figure 12.8 shows the results graphically.

Table 12.11 Grammatical ability by frequency of use of English

Frequency of use	Mean	N
Cannot say	.649	103
Once a month or less often	.610	142
Once every two weeks	.646	94
Once or twice a week	.702	474
(Almost) every day	.936	271

Self-assessments of grammatical ability

Since there are no self-assessment statements for Grammar, it is not possible to analyse the correspondence between self-assessed grammatical ability and test-based grammatical performance. However, it is possible to conduct an analysis of the draft self-assessment statements for grammatical ability that were developed for the standard-setting procedure, as reported in Chapter 6 and above.

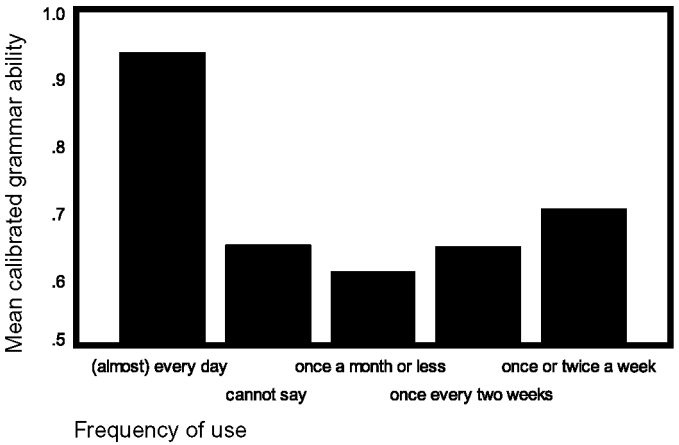

Figure 12.8 Grammatical ability by frequency of use of English

The rank order of difficulty of grammatical structures, as perceived by the standard-setting judges, was as shown in Table 12.12.

The correlation between the supposed CEFR level, as intended by the DIALANG levels, and the mean perceived CEFR level of the grammatical descriptors (in the opinion of ten judges) was .98 for the 32 descriptors, and the correlation between the intended level and the median level was also a high .958. The crosstabulation was as shown in Table 12.13 (p. 188) (Chi square = 105.568, df 25, p = .000).

Clearly, there is considerable overlap, with most uncertainty occurring at B1 and C1. We can thus begin to have some confidence in the hypothesized levels of the various descriptors, and this is a reasonably solid basis for further research into the development of grammatical ability.

Similar data are also available for other DIALANG languages, and Table 12.14 (p. 188) shows the correlations between intended CEFR level and the average judged level of the descriptors from the standard-setting procedures for a number of DIALANG languages.

This amount of agreement is entirely consistent with and very similar to agreement on descriptors for Reading, Listening and Writing, for English, German, Spanish and French.

Relationship between Grammar and other macro skills

In order to examine the relationship between the macro skill of Grammar and the other two skills which had been tested alongside Grammar, correlations were run, as shown in Table 12.15 (p. 188).

The correlations are significant and substantial, showing there is indeed a quite close relationship between grammatical ability and both Reading and Writing. Given this close relationship, it was decided to

Table 12.12 Order of perceived difficulty of grammatical descriptors

Descriptor number	Descriptor wording	Median level	Mean level (A1 = 1, A2 = 2, etc.)
1.1	Learners show limited and unsystematic control of a few simple grammatical structures and sentence patterns.	A1	1.10
1.2	Learners can form simple main clauses but not necessarily accurately.	A1	1.20
1.3	Learners know the personal pronouns in their basic form.	A1	1.30
1.5	Learners can respond to basic formulaic questions, e.g. *Who are you? Where do you come from? Do you speak English?*	A1	1.30
1.4	Learners know the most frequent question words and can produce some formulaic questions.	A2	1.60
2.1	Learners use some simple structures correctly.	A2	1.80
2.4	Learners can formulate simple sentences and combine them to some degree by using the most frequent conjunctions, e.g. *and, or, because*.	A2	2.10
2.2	Learners can use the most frequent verbs in the basic tenses (present and simple past).	A2	2.10
2.3	Learners can use verb forms in other than the basic tenses only in fixed phrases, e.g. *I would like to have coffee*.	A2	2.10
2.5	Learners can form simple questions and negations with the most common auxiliary verbs, e.g. *can*.	A2	2.20
2.6	Learners can use personal pronouns more or less systematically to distinguish between singular and plural and between simple cases, e.g. *them vs their*.	B1	2.50

Table 12.12 (continued)

Descriptor number	Descriptor wording	Median level	Mean level (A1 = 1, A2 = 2, etc)
3.6	Learners know the basic word order well.	B1	3.10
3.4	Learners know the comparison of adjectives.	B1	3.20
3.2	Learners know the past tenses and can produce both their positive and negative forms.	B1	3.30
3.3	Learners know the basic principles of constructing passives.	B2	3.50
3.5	Learners can formulate the most frequent expressions and clause types accurately; they can use, e.g., some conjunctions, auxiliary verbs and frequent adverbs systematically.	B2	3.60
4.5	Learners can use pronouns quite systematically. Learners know the comparison of adverbs.	B2	3.60
3.1	Learners have a repertoire of frequently used grammatical patterns, which they can use reasonably accurately.	B2	3.70
4.3	Learners have a good command of most adverbs.	B2	3.80
4.4	Learners can use articles for the most part appropriately.	B2	3.90
4.6	Learners show some ability to vary their usage through paraphrases.	B2	4.00
4.1	Learners have a good control of all basic grammatical structures and they can sometimes produce complex structures accurately, i.e. they have some control of the passive voice, conditionals and modal verbs.	B2	4.10
4.2	Learners can use conjunctions in order to express causal and conditional relationships.	B2	4.20

Table 12.12 (continued)

Descriptor number	Descriptor wording	Median level	Mean level (A1 = 1, A2 = 2, etc)
5.4	Learners can use articles in many less frequent contexts.	B2	4.30
5.2	Learners know the standard written language conventions, e.g. punctuation marks (including commas) and abbreviations.	B2	4.30
5.3	Learners can use many phrasal verbs and prepositional expressions naturally.	C1	4.70
5.1	Learners can produce complex expressions, such as relative clauses, naturally and accurately.	C1	5.00
5.5	Learners may make occasional mistakes in rare or complex structures.	C1	5.30
6.3	Learners can vary their usage through paraphrases flexibly and naturally.	C2	5.50
6.4	Learners have a very good command of the conventions of written language.	C2	5.60
6.1	Learners show an accurate control of a very broad range of grammatical structures; they can use, e.g., passive voice, inversions, and complex conditionals.	C2	5.70
6.2	Learners can choose structures that fit the topic and style/genre of the text very well.	C2	5.90

examine the contribution of the various subskills of Grammar to the two macro skills, in comparison with their contribution to Grammar. Results are shown in Table 12.16 (p. 189).

Table 12.16 shows that the contribution of each of the subskills to an overall measure is only moderate, although some correlations are more substantial than others. However, this is also the case for the macro skill of Grammar, not just for Reading and Writing, and this is almost

Table 12.13 Crosstabulation between intended and judged level of descriptors, for Grammatical Structures

		Median rated level						Total
		A1	**A2**	**B1**	**B2**	**C1**	**C2**	
DIALANG level	**A1**	4	1	0	0	0	0	5
	A2	0	5	1	0	0	0	6
	B1	0	0	3	3	0	0	6
	B2	0	0	0	6	0	0	6
	C1	0	0	0	2	3	0	5
	C2	0	0	0	0	0	4	4
Total		4	6	4	11	3	4	32

Table 12.14 Correlation between intended and perceived CEFR levels of grammatical descriptors

Language	Mean with CEFR level	Median with CEFR level	Number of judges	Number of descriptors
German	.987	.988	7	40
Spanish	.960	.941	12	18
French	.975	.971	7	17

certainly due in part to the small number of items testing any one subskill, and possibly to the somewhat disparate nature of what is being tested even within any one subskill. Modest correlations are seen with Writing, despite the fact that it might be argued that many indirect Writing items are actually disguised tests of grammar. However, we would need many more items testing each subskill before we could have confidence in any conclusions that could be drawn from such correlations.

In an attempt to understand better the contribution of grammatical subskills to macro skills, a number of factor analyses were run. However, in no case could a factor analysis be completed, either because there were fewer than two cases, or because at least one of the variables had zero variance, there was only one variable in the analysis, or correlation coefficients could not be computed for all pairs of variables.

Table 12.15 Relationship between Grammar, Reading and Writing

	Reading ability	Writing ability
Grammar ability	.685	.769
N	718	366

Table 12.16 Grammar subskills and Grammar, Reading and Writing

	Grammar ability	Reading ability	Writing ability
Inflections of adj and adv (k = 1)	NS	.176	NS
Proper and common nouns (k = 4)	.618	.440	.433
Verbs active and passive (k = 4)	.543	.377	.601
Syntax punctuation (k = 4)	.344	.261	.227
Morphology – numerals (k = 8)	.505	.345	.481
Morphology other (k = 6)	.470	.447	.344
Syntax organization (k = 10)	.652	.401	.576
Pronouns – context (k = 6)	.521	.394	.387
Verbs, tense, mood, etc. (k = 6)	.592	.419	.406
Comparison of adj and adv (k = 4)	.423	.355	.273
Definite/indefinite articles (k = 2)	.371	.240	.361
Simple and complex sentences (k = 1)	.264	.180	.204

No doubt the basic reason for this was the small number of items in any one subskill and the fact that not every subskill was contained in each pilot test booklet. Before useful conclusions can be drawn about grammatical ability, the subskills will have to be combined in some way, or there will need to be many more items created for each subskill, or research will need to be conducted with a full factorial design for each subskill.

Summary and conclusion

In this chapter, we have discussed the construct of grammar, described how DIALANG decided to operationalize that construct, and examined a number of test items contained in the English Grammar test in Version 1. The performance of Grammar items was examined in relation to the CEFR and it was shown that there was no reason to believe that as items become more difficult, they tend to test certain subskills and not others.

When the abilities of learners taking the Grammar test were examined in relation to background variables, it was found that there were significant differences in grammatical ability by mother tongue, sex, educational level, length of time learning English and frequency of use of English, but not by age.

Although grammatical abilities were shown to have a reasonably close relationship to the macro skills of Reading and Writing, the subskills that were tested in DIALANG proved not to be closely related to these macro skills. It was suggested that this was likely to be due to the small number of items testing any one subskill, and to the design of the pilot test booklets. In order to explore the relationship between components of grammatical ability and the various skills, it will be

important to have a larger number of items testing each subskill, to define the subskills quite carefully to ensure homogeneity within the subskill, and to devise a research design that ensures that each learner takes tests of each subskill, as well as of each macro skill. This is likely to involve a fairly extensive piece of research, but the results should be of considerable interest, since it is quite likely that diagnosis of the macro skills could be notably enhanced by developing tests of grammatical abilities that can be shown to predict strengths and weaknesses in the macro skills. No doubt a close examination of the results of research in second language acquisition, for English at least but hopefully also for languages like German, Spanish and French where not inconsiderable research has been conducted, might help inform the development of relevant measures. We will return to this topic in the final chapter.

Chapter 13: Vocabulary

I begin this chapter by discussing what is meant by the construct of 'vocabulary', then describe DIALANG's approach to the assessment of vocabulary (separate from the assessment of vocabulary size in the Vocabulary Size Placement Test described in Chapter 7). I exemplify the construct with a number of items taken from Version 1. I then analyse the performance of the piloted items, and look at how the scores of learners taking pilot Vocabulary tests related to background variables. Finally I examine the relationship between vocabulary subskills and macro skills of Vocabulary, Listening and Writing, and draw tentative conclusions for further research and development.

The construct

For a long time, foreign language assessment was much more concerned with the assessment of grammar, and of the classical four skills, than it was with testing a learner's vocabulary. Traditional discrete-point tests of vocabulary had indeed existed in the 1960s and 1970s, but these were largely discredited during the 'communicative revolution' alluded to in Chapter 12. The testing of individual lexical items was felt to be unnecessary and misleading: unnecessary because it was argued that one was testing a learner's vocabulary abilities when assessing their ability to understand written and spoken texts as well as their ability to produce text in writing and in speech; and misleading because most vocabulary tests simply could not be long enough to cover an adequate sample of the words in a language and therefore would be biased by the – more or less arbitrary – selection of words sampled by the test.

More recently, however, there has been a resurgence of interest in the assessment of vocabulary, and indeed in the role of vocabulary knowledge in foreign language performance and proficiency more generally. Read (2000) is an excellent example of this renewed concern with the assessment of vocabulary, and he provides a very useful overview of what is involved in assessing a learner's vocabulary and lexical abilities. After discussing such basic questions as 'What is a word?', 'How

important are word combinations (or lexical phrases, as they have come to be called)?', 'What does it mean to know a lexical item?', 'What is the nature of vocabulary ability?', Read reviews recent trends in vocabulary assessment and research, and distinguishes three different dimensions in vocabulary assessment: discrete–embedded; selective–comprehensive; and context-independent–context-dependent. These three dimensions inform his discussion of vocabulary research, and his presentation and exemplification of procedures for assessing vocabulary.

DIALANG is, in a sense, a product of this renewed interest in vocabulary, not only because it includes a Vocabulary component in its battery of tests, but also because it uses a Vocabulary Size Test as a placement procedure, and is thus able to provide learners with quite a lot of feedback on their vocabulary abilities. For more information on the Vocabulary Size Placement Test and its value in DIALANG, see Chapter 7.

Although it is difficult for a computer-scored test to measure a learner's ability to use appropriate vocabulary accurately in context, DIALANG nevertheless attempts, through the use of different item types, to measure a learner's productive as well as receptive vocabulary knowledge. And since the Vocabulary test is intended to be diagnostic, it is not seen as problematic that the test items are based upon relatively short stretches of context, since the aim is more to explore the nature of a learner's lexical abilities than it is to know how well the learner can derive the meaning of unknown words from context. This latter subskill is instead tested as part of the Reading test.

The DIALANG Assessment Specifications (DAS) describe in some detail the construct that the Vocabulary test is intended to measure. These specifications are only in part derived from the CEFR, because the CEFR is neutral as to which language is to be taught or tested, and therefore has little to say about the nature of vocabulary in particular languages, or about the nature of lexical ability. DIALANG therefore had to turn to other sources for insight into what sort of useful diagnostic information could be derived from assessment.

Despite the recent interest in vocabulary acquisition, there is still not much empirical evidence regarding how foreign language learners' lexical competence develops, and so DIALANG recognized the impossibility of developing specifications for each CEFR-related level separately. Instead, the DAS merely indicated to item writers the sorts of things they might take into consideration when writing test items aimed at the various CEFR levels. As we will see, item writers were asked to indicate what level was intended by each item, but this did not necessarily bear any close relationship to the empirical difficulties that resulted from the piloting.

Nevertheless, DIALANG was quite specific in stating that it was crucial for the success of diagnosis that all test items should be piloted, their difficulties calibrated and some means be found to relate the items to the

CEFR *post-hoc*. This was accomplished, as we saw in Chapter 6, through standard-setting procedures.

In addition, as part of the development of CEFR-related standard-setting procedures for Vocabulary, DIALANG began to develop a set of self-assessment statements which might eventually be used both to inform vocabulary test specifications and to contribute to the development of self-assessment statements (as described in Chapter 6). Towards the end of this chapter, we will examine the empirical results of the standard-setting procedures, to see what light they might be able to throw on the development of vocabulary knowledge and lexical ability.

The construct of the DIALANG Vocabulary test

The Vocabulary test aims at assessing the learner's knowledge of different aspects of words, focusing on the mastery of the meanings of simple word units and combinations. In creating the tasks, four dimensions of word meaning were distinguished: denotative meaning, semantic relations, combinations and word formation. These were defined as the ability to:

1 recognize/produce word meanings, including denotation, semantic fields, connotation, appropriateness;
2 recognize/produce semantic relationships between words, including synonymy/antonymy/converses, hyponymy/hypernymy, polysemy;
3 recognize/produce word combinations including collocation and idiomaticity;
4 recognize/produce words by compounding and affixation.

Sampling words to be tested

The Test Development Teams chose between two systematic frames for sampling items. The Project adopted a sampling approach rather than starting from the teams' perception of the six Council of Europe proficiency levels because detailed descriptors for vocabulary levels have not been published in the Council of Europe literature. The Project felt that random sampling would offer a secure basis for constructing an initial item pool, and the level assignment could be done through piloting.

In order to use a sampling approach, a basic corpus had to be used for each language from which the sampling could be made. The preference would have been to use a frequency word list for all languages, but recent versions of such lists did not exist for all the DIALANG languages. Therefore, two slightly different sampling frames were created. The first was based on frequency lists and the second on dictionary sampling. The teams used the frequency-based method if a relatively recent frequency list was available, and dictionary sampling if it was not. The teams were asked to take care that base words and derivates were counted as different

words, while inflections were not counted as separate words. Detailed instructions were given for each sampling procedure.

Language-specific modifications

The teams were asked to write 1,500 vocabulary items, but because the items were sampled from a corpus rather than selected to fit the above categorization, the teams were not asked to produce an equal number of items for all the categories.

The teams were also asked to estimate the difficulty of their items. They based their estimates on their experience of teaching their language to learners, and on the frequency estimation of the item in the corpus.

Test items

In this section, we exemplify the various item types used in DIALANG, by showing sample items from the English Vocabulary test in Version 1, in the screenshots in Figures 13.1–13.4.

Analysis of the results of piloting of the English Vocabulary Test

As a result of piloting, all 60 Vocabulary items survived into Version 1. Table 13.1 shows the number of each item type in Version 1, and their mean empirical difficulty.

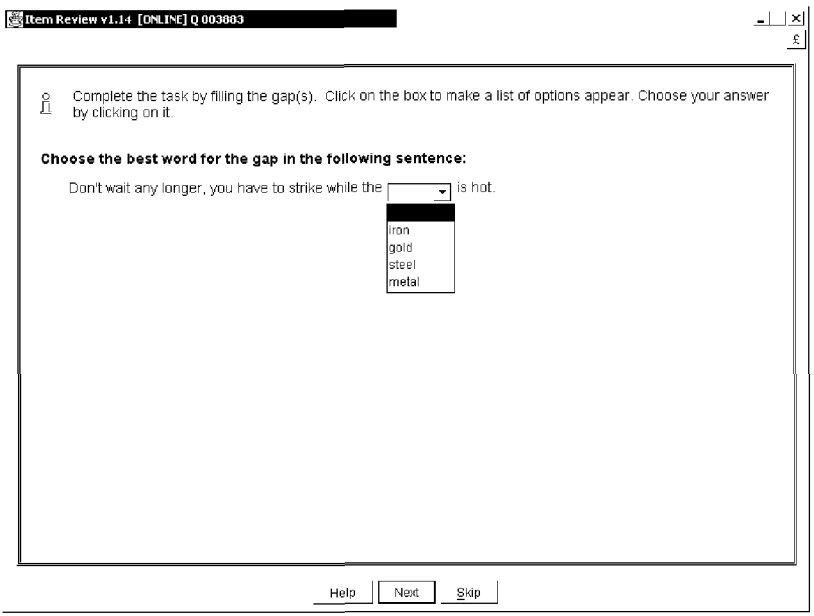

Figure 13.1 Combination

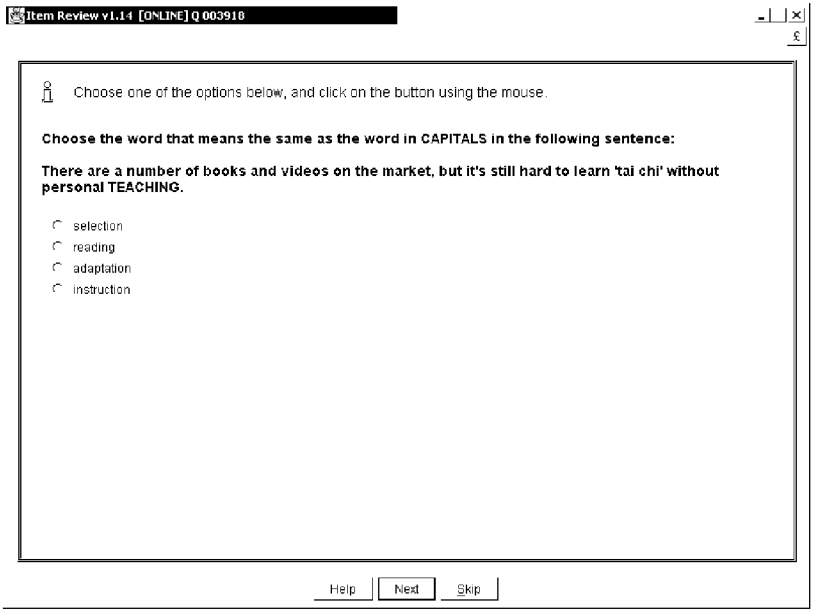

Figure 13.2 Semantics

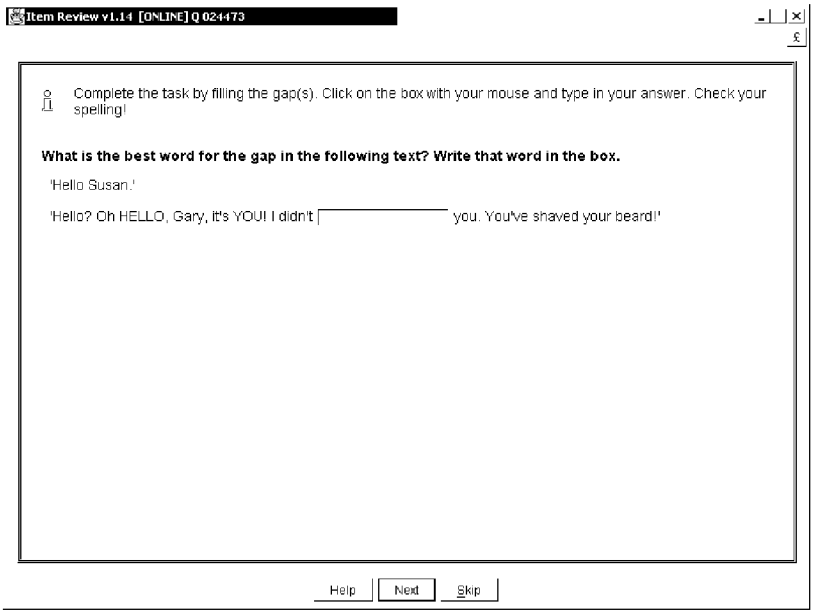

Figure 13.3 Meaning

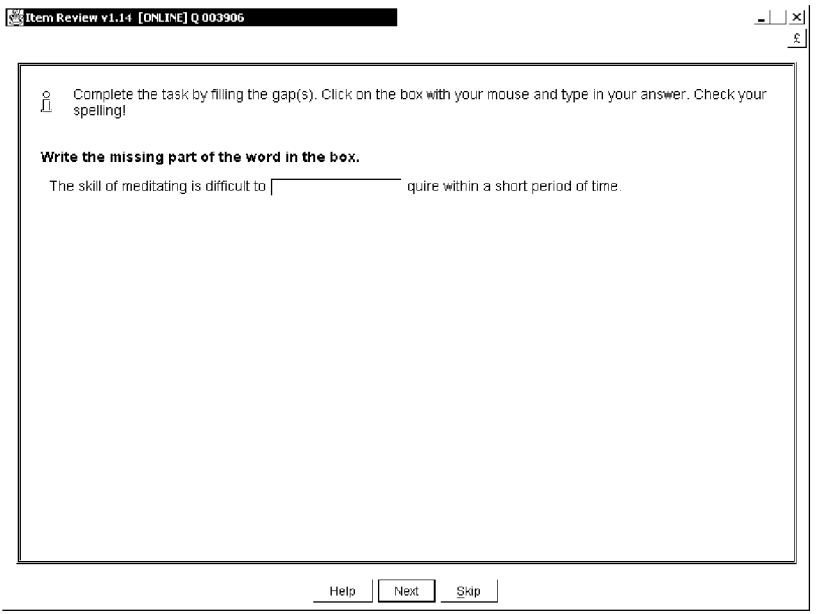

Figure 13.4 Word formation

Table 13.1 Difficulty by item type

Item type	N	Mean	Min	Max	Std Deviation	
Drop-down	19	−.158	−1.272	1.872	.6751	
Multiple-choice	12	−.076	−.621	1.103	.5352	
Short-answer	2	.011	−.275	.296	.4038	
Text-entry	27	.144	−.902	.848	.4332	F = 1.254 P = .299

A one-way ANOVA revealed no significant differences in mean scores by item type. Thus although the trend seems to suggest that productive item types (short-answer and text-entry) are more difficult than selected-response types, statistically there was no difference.

Despite the highly significant difference among ABG levels, the only significant contrast was between A and C levels (as intended by the item writers). See Table 13.2.

If we classify the items according to their empirically derived CEFR levels, and then crosstab those levels with the ABG levels, we find no significant association ($\chi^2 = 29.726$, p = .194).

Table 13.3 shows the four subskills and their mean difficulties. A one-way ANOVA revealed no significant difference in mean scores across subskills.

Table 13.2 Difficulty by Author's Best Guess

Level	N	Mean	Min	Max	Std Deviation		
A	21	−.322	−1.272	.423	.4643		
B	20	.002	−.667	.848	.3509		
C	19	.354	−.613	1.872	.5908	F = 10.061	P < .000

Table 13.3 Difficulty by subskills

Subskill	N	Mean	Min	Max	Std Deviation		
Combination	13	.070	−1.272	1.872	.7612		
Meaning	19	−.023	−.774	.675	.4807		
Semantic relations	16	−.083	−.948	1.103	.5020		
Word formation	12	.071	−.902	.848	.4632	F = .267	P = .849

If we take the CEFR level of each item, and then crosstab CEFR level by subskill, we find no significant association ($\chi^2 = 7.172$, $p = .846$). Thus we can safely conclude that there is no difference in difficulty across items by the subskill they test.

Vocabulary ability

So far we have examined the Vocabulary items and their performance during piloting. In this section, we examine the scores of those learners who took English Vocabulary tests as part of the piloting. Those who took the pilot tests were assigned at random to a test booklet, two of which contained Writing and Vocabulary tests, and four of which contained Listening and Vocabulary tests. This means that we can compare learners' performance on Vocabulary tests with their performance on either Writing tests or Listening tests.

Before I present an analysis of results on the Vocabulary test according to background variables and in relation to other skills, I present the descriptive statistics for scores on the Vocabulary test (Table 13.4). As in other chapters, scores were calibrated using IRT and are

Table 13.4 Descriptive statistics for vocabulary ability

N	Minimum	Maximum	Mean	SE Mean	Std Deviation
975	−1.460	2.480	.482	.0174	.5448

expressed in logit units, which range from a minus figure (for a relatively weak performance) to positive figures for stronger performances.

Vocabulary knowledge by mother tongue

Results are presented in Table 13.5. A one-way ANOVA revealed significant differences overall ($F = 12.317$, $p = .000$). Paired comparisons showed considerable overlap, but with some significant contrasts for Finnish, French, Portuguese, Spanish and Other L1s.

Table 13.5 Vocabulary knowledge by mother tongue

Mother tongue	Mean	N
Icelandic	.795	40
Danish	.712	58
German	.686	223
Norwegian	.668	43
Swedish	.611	20
Dutch	.494	114
Italian	.398	14
Finnish	.373	160
French	.315	69
Other L1s	.257	159
Portuguese	.236	19
Spanish	.213	55

Table 13.6 Vocabulary knowledge by sex

Sex	Mean	N
Female	.509	598
Male	.438	377

Vocabulary knowledge by sex

See Table 13.6. Females just score significantly higher than males ($p = .046$).

Vocabulary knowledge by age

Table 13.7 shows the data. Interestingly, this time vocabulary knowledge does indeed vary by age, with older users scoring higher ($F = 6.108$) and with under 18s scoring significantly lower than older groups. Figure 13.5 shows this result graphically.

Table 13.7 Vocabulary knowledge by age

Age	Mean	N
Under 18	.170	21
18–25	.450	600
26–35	.450	189
36–45	.661	99
46–55	.645	61
56–65	1.142	3
Over 65	1.372	2

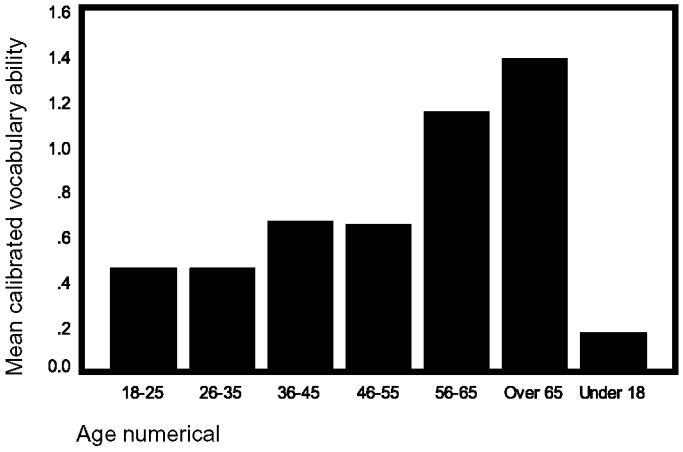

Figure 13.5 Lexical ability by age

Vocabulary knowledge by educational level

See Table 13.8. A one-way ANOVA revealed significant differences across educational levels ($F = 15.204$, $p = .000$), with university level significantly higher than all other groups.

Table 13.8 Vocabulary knowledge by educational level

Education	Mean	N
Higher (university)	.593	510
Other	.560	20
Secondary (general)	.429	191
Higher (non-university)	.413	143
Secondary (vocational)	.144	68
Primary	.121	43

Vocabulary knowledge by length of time learning English

Results are presented in Table 13.9. A one-way ANOVA showed significant differences across length of time learning English ($F = 32.015$), with longer periods scoring higher. Figure 13.6 shows the results graphically.

Table 13.9 Vocabulary knowledge by length of time learning English

Length of time	Mean	N
Less than a year	−.120	26
1–2 years	−.035	51
3–4 years	.262	99
5–6 years	.395	201
7–8 years	.508	189
9–10 years	.565	202
More than 10 years	.769	207

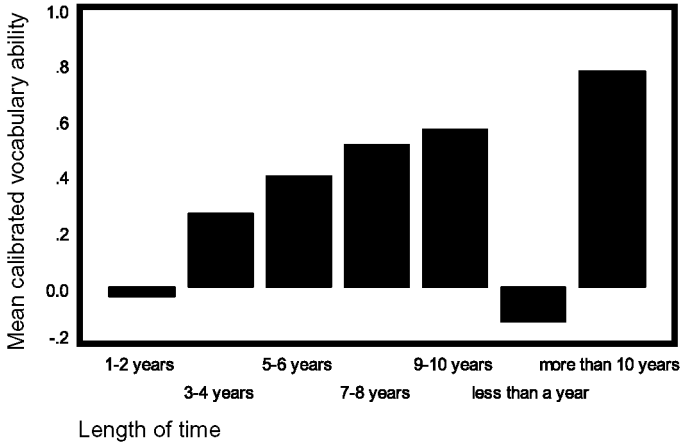

Figure 13.6 Lexical ability by length of time learning English

Vocabulary knowledge and frequency of use of English

See Table 13.10. A one-way ANOVA showed significant differences by frequency of use ($F = 30.732$, $p = .000$), with more frequent use associated with much higher scores. Figure 13.7 shows these results graphically.

Table 13.10 Vocabulary knowledge and frequency of use of English

Frequency of use	Mean	N
Cannot say	.279	87
Once a month or less often	.378	122
Once every two weeks	.300	82
Once or twice a week	.412	433
(Almost) every day	.783	251

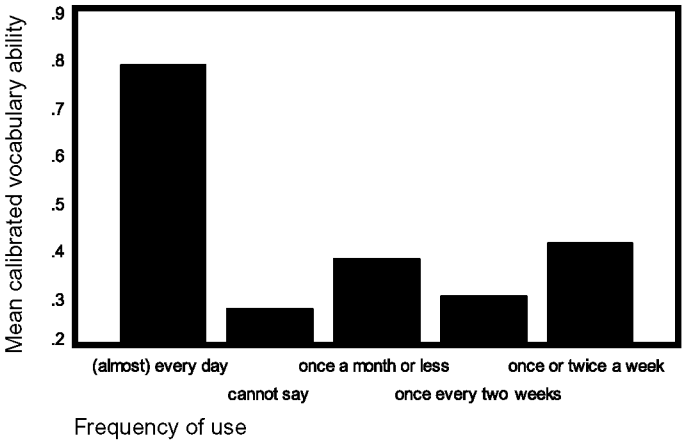

Figure 13.7 Lexical ability by frequency of use of English

Self-assessment of lexical abilities

Since there are no self-assessment statements for Vocabulary, it is not possible to analyse the correspondence between self-assessed lexical ability and test-based Vocabulary performance. However, it is possible to conduct an analysis of the draft self-assessment statements for lexical ability that were developed for the standard-setting procedure, as reported in Chapter 6 and above.

The rank order of difficulty of lexical descriptors, as perceived by standard-setting judges, was as shown in Table 13.11. It is worth noting that several statements intended to be at a particular level turned out to be at higher or lower levels (for example, V3.4, V2.1, V4.4, V3.3, V5.1).

The correlation between the supposed CEFR level, as intended by the DIALANG levels, and the mean perceived CEFR level of the lexical descriptors (in the opinion of nine judges) was .947 for the

Table 13.11 Order of perceived difficulty of lexical descriptors

Descriptor number	Descriptor wording	Median level	Mean level (A1 = 1, A2 = 2, etc.)
V1.2	Learners can use some words and phrases meaningfully in everyday contexts.	A1	1.22
V1.3	Learners can recognize the opposites of the most common words, e.g. *good – bad, man – woman*.	A1	1.22
V1.4	Learners can recognize the meaning of the most simple derived words, e.g. *teacher, shopper*.	A1	1.44
V1.1	Learners know the approximate meaning of around 500 words used in the most common contexts.	A1	1.78
V2.2	Learners know the opposites of very frequent words, and can recognize synonyms for some words.	A2	2.00
V3.4	Learners can use synonyms of some very common words, e.g. *nice – kind* (person).	A2	2.33
V2.3	Learners recognize some basic principles of word formation, and can also apply some of them, e.g. *to drive – a driver*.	A2	2.33
V3.5	Learners can use some frequent collocations, e.g. *a tall girl – a high mountain*.	B1	2.67
V2.1	Learners know and can use everyday vocabulary related to a range of basic personal and familiar situations.	B1	2.67
V3.2	Learners can use a range of prefixes to produce opposites to basic words.	B1	3.22
V4.4	Learners recognize and know how to use the basic word formation principles, e.g. *note – notify; fur – furry; accident – accidental; paint – painting*.	B1	3.56
V3.1	Learners have a good command of vocabulary related to everyday situations.	B2	3.67

Table 13.11 (continued)

Descriptor number	Descriptor wording	Median level	Mean level (A1 = 1, A2 = 2, etc.)
V4.2	Learners can recognize words with more than one meaning, e.g. *back a car/back a proposal*.	B2	3.89
V4.3	Learners know a number of frequently used idioms.	B2	3.89
V4.1	Learners can use the synonyms and opposites of many common words in different contexts.	B2	4.00
V3.3	Learners know a number of principles of word formation, e.g. *to agree – agreeable; a verification – to verify*.	B2	4.11
V4.5	Learners can express meanings by adding prefixes and affixes to familiar words, e.g. *to re-send a message; to have an after-dinner nap*.	B2	4.11
V5.4	Learners know word formation not only for normal words but also for less common expressions.	C1	5.00
V5.2	Learners know the synonyms and opposites of less common words, e.g. know how to use *vast, abundant* and *scant*.	C1	5.33
V5.3	Learners know many collocations for a range of contexts, and can choose words in a context-sensitive way.	C1	5.33
V6.2	Learners can combine words idiomatically and know their collocations also in less frequent contexts.	C2	5.67
V5.1	Learners can vary expressions idiomatically by using parallel phrases, synonyms and opposites.	C2	5.89
V6.1	Learners can understand and use abstract, figurative, idiomatic and colloquial language as appropriate to the situation.	C2	6.00

Table 13.12 Crosstabulation between intended and judged level of descriptors, for Lexical Descriptors

		Median						Total
		A1	**A2**	**B1**	**B2**	**C1**	**C2**	
DIALANG	**A1**	4	0	0	0	0	0	4
	A2	0	2	1	0	0	0	3
	B1	0	1	2	2	0	0	5
	B2	0	0	1	4	0	0	5
	C1	0	0	0	0	3	1	4
	C2	0	0	0	0	0	2	2
Total		4	3	4	6	3	3	23

Table 13.13 Correlation between intended and perceived CEF levels of lexical descriptors

Language	**Mean with CEF level**	**Median with CEF level**	**Number of judges**	**Number of descriptors**
German	.983	1.00	7	27
Spanish	.977	.988	12	20
French	.968	.957	7	13

23 descriptors, and the correlation between the intended level and the median level was also a high .945. The crosstabulation was as shown in Table 13.12 (Chi square = 69.256, df 25, p = .000).

Clearly, there is considerable overlap, with what uncertainty there is occurring at B1. We can thus begin to have some confidence in the hypothesized levels of the various descriptors, and this is a reasonably solid basis for further research into the development of lexical ability.

Similar data are also available for other DIALANG languages, and Table 13.13 shows the correlations between intended CEFR level and the average judged level of the descriptors from the standard-setting procedures for a number of DIALANG languages.

This amount of agreement is entirely consistent with and very similar to agreement on descriptors for Reading, Listening and Writing, for English, German, Spanish and French.

Relationship between subskills and various macro skills

It will be recalled that those who took a Vocabulary test also took either a Writing test or a Listening test. This enables us to examine the

Table 13.14 Intercorrelations of macro skills

	Writing ability	Listening ability
Vocabulary ability	.795	.650
N	369	606

Table 13.15 Vocabulary subskills with Listening and Writing

	Writing ability	Listening ability
Vocab combination	.628	.431
Vocab word formation	.706	.502
Meaning	.656	.560
Semantic relations	.619	.443

relationship among these skills and to explore how the subskills of Vocabulary relate to the macro skills.

The rank order correlations among the three macro skills were as in Table 13.14.

Clearly vocabulary knowledge plays a substantial role in both writing and listening. To explore this relationship further, Table 13.15 reports the intercorrelation of Vocabulary subskills and the two macro skills of Writing and Listening.

Although Vocabulary shows a closer relationship to Writing, this is doubtless in part because many of the indirect Writing items are, as we have seen, rather similar to some of the Vocabulary items. It is interesting that the individual Vocabulary subskills also bear a reasonably close relationship to the macro skill of Listening, and in particular the subskill of meaning.

Factor analyses sought to throw more light on these relationships, but in every case, only one factor emerged from the analyses, accounting for between 63 per cent and 73 per cent of the variance. Nevertheless, it is worth noting the loading of the subskills on the single factor for writing

Table 13.16 Vocabulary subskills and Writing

	Factor 1
Vocab combination	.736
Vocab word formation	.807
Meaning	.756
Semantic relations	.736
Writing ability	.861

Table 13.17 Vocabulary subskills and Listening

	Factor 1
Vocab combination	.563
Vocab word formation	.710
Meaning	.794
Semantic relations	.663
Listening ability	.706

ability, and separately for listening ability, as reported in Tables 13.16 and 13.17.

Clearly there are reasonably close relationships between the subskills of Vocabulary and the macro skills of Writing and Listening. This suggests that diagnostic tests of the subskills of Vocabulary might prove to be rather useful in diagnosing strengths and weaknesses in Writing and Listening.

Summary and conclusion

In this chapter, we have discussed the construct of vocabulary ability, described the construct that DIALANG decided to test and illustrated this with a number of items from Version 1 of the English Vocabulary test. We then examined the performance of the English Vocabulary items, establishing that there was no reason to believe that the different CEFR levels could easily be characterized as involving different vocabulary subskills. Conversely, one could conclude that it is possible to test the various vocabulary subskills at each CEFR level: there is no evidence from these analyses that as vocabulary ability develops, different vocabulary subskills are engaged.

An analysis of performance on the Vocabulary test by learners' background variables showed significant differences by mother tongue, sex, age, level of education, length of time learning English, and, to some extent, frequency of use of English, although in the latter case the most marked advantage was to those who used the language almost every day, and lower frequencies of use did not result in greatly enhanced performance.

Vocabulary ability was seen to relate quite closely both to writing ability and to listening ability. It was suggested that this implies that there might be merit in exploring further the potential of tests of aspects of vocabulary for the diagnosis of strengths and weaknesses in these macro skills. Conceivably, although the piloting design did not allow an exploration of this hypothesis, it may be the case that diagnostic tests of vocabulary might also be useful in predicting performance in reading. Further research is clearly called for to examine this possibility.

Furthermore, and despite the lack of a theory of the development of language skills in general, or of vocabulary ability in particular, it may very well be the case that development in the three macro skills is parallel to, and conceivably caused by, development in a learner's vocabulary, both in terms of the size of the vocabulary and the quality of the learner's understanding and use of vocabulary. These are issues to which we will return in the final chapter.

Chapter 14: The value of feedback and advice

A crucial component of any diagnostic test must be the feedback that is offered to users on their performance. Merely presenting users with a test score, without any acompanying explanations as to what that score might mean, is clearly not very helpful, even on proficiency or achievement tests, but it is quite inappropriate on diagnostic tests. The essence of a diagnostic test must be to provide meaningful information to users which they can understand and upon which they or their teachers can act. Moreover, that feedback must be given as soon as possible after the user has taken the test. There is very little value in feedback that comes two or three weeks, or even months, after one has taken the test, since inevitably one will have forgotten how one performed and why one responded the way one did.

As pointed out in Chapter 1, with the advent of computer-based tests, the provision of immediate feedback to test users is a relatively simple matter, thereby making it potentially highly informative and relevant. It is probably no accident that diagnostic testing became a more interesting possibility once computer-based testing had become both feasible and cheap.

In this chapter we discuss the feedback provided by DIALANG, because it is one of the most developed systems available at present.

The feedback component of DIALANG is central to diagnosis. But not only that. As we have seen, part of the DIALANG philosophy is that learners should be responsible for their own learning, especially in settings where they may not be taking regular, institutionalized language lessons. Indeed, the importance of autonomy in language learning has been well recognized for decades, and has been emphasized in particular by the Council of Europe, through its many publications on the topic (e.g. Holec, 1980, 1992, 1994; Riley *et al.*, 1996) as well as by others (Brookes and Grundy, 1988; Dickinson, 1987; Gathercole, 1992; Little, 1991, and others). The feedback in DIALANG is thus aimed at informing learners, at supporting learning and at raising awareness, both about one's own performance and abilities and about what is involved in language learning, and how one's language level might be enhanced.

We saw in Chapter 8 how self-assessment is thought to contribute to autonomous language learning, by giving learners more control over their language learning, by enhancing their self-awareness and their awareness of the language learning process, and by giving them a realistic idea of their own abilities, since this is believed to lead to more successful language learning in the long term. An accurate estimation of one's own language ability is believed to be important for setting an appropriate goal and identifying potential weaknesses in the process of one's language learning, especially for self-directed adult learners.

Similarly with feedback: since DIALANG is diagnostic, it aims to inform and support learners through the feedback it gives, which is intended to be maximally informative and useful. Thus self-assessment, the provision of feedback and the encouraging of learner reflection go hand-in-hand.

In this chapter, we will explore to what extent the feedback in DIALANG meets these goals.

Chapter 3 described in some detail what sort of feedback DIALANG users are provided with on their test performance and its relationship with self-assessment. Users are free to look at all or none of this feedback, in keeping with the DIALANG philosophy that everything in DIALANG is optional – users can even skip out of a test if they prefer not to complete it, but if they do, they will only receive feedback on those items they have completed before quitting.

To summarize briefly the range of feedback available, users can receive their test result in terms of a CEFR level, and a brief verbal explanation of what that level means, again based on the CEFR. They can see which items they got right and which were wrong, grouped according to the subskill being tested by each item. They can also click on any test item in the display and see what their response was as well as what the correct response was. This information is also available to users during the test if they select immediate feedback mode.

Users can choose to see how their self-assessment compared with their test result, and a separate section of the feedback allows them to explore possible reasons why there was a mismatch (if any) between the two. Feedback is also available on their Vocabulary Size Placement Test performance, in the form of one of six score bands, with brief descriptors explaining what the score means.

One major section of feedback is what is called in DIALANG 'Advisory Feedback'. Users are presented with a set of tables that explain some of the differences between what a learner can do at the level to which the user was assigned, and what learners at lower and higher levels can do. They explain what types of text a learner at a given level can cope with, what they can understand, and under what conditions and limitations. These tables are based on the CEFR (and are contained in Appendix C of the 2001 edition of the Framework – pp. 238–43 in

the English edition). Users can also explore advice on how they can improve from their current level to higher levels.

Finally, it must be remembered that all this feedback, as well as the self-assessment, is available, not only in the target language, but above all in the ALS, the language of administration. This means that users, at least those whose mother tongue is one of the DIALANG languages, can understand and appreciate the feedback and advice they are given.

An important feature of any feedback must be that it not only relates to what a learner can recognize as being relevant, but that it can also relate to whatever materials, curricula or opportunities there are in the learners' world for improving their language ability. This means that diagnostic tests and the feedback they enable and provide should be closely related to learning opportunities. This is one of the major reasons why the DIALANG system was based on the Common European Framework. As explained in Chapter 4, the CEFR was at the time the only pan-European framework available that could conceivably link testing to teaching and learning. Other frameworks existed, but were either foreign in origin and bias (North American, Australian, Hong Kong Chinese) or had their origins in only one European country (e.g. the UK's National Curriculum and its associated Attainment Targets and Assessment Tasks).

But the greatest advantage of the CEFR was not simply that it existed, nor even that it was European in the broadest sense, but above all that it was specifically intended to integrate assessment, learning and teaching by being a common point of reference for all. Thus any test based on the CEFR would, at least potentially, be directly relevant to curricula, textbooks and other learning materials which were also based on the CEFR. And with the advent of the European Language Portfolio, also based on the CEFR, DIALANG would also complement alternative methods of recording and assessing progress in language learning. Thus diagnosis could be directly related to teaching and learning.

The development of feedback and advice

In Phase 1, in parallel to the development of the tests and the software, a project team was set up to develop both the self-assessment and the feedback and advice to be presented in DIALANG – since self-assessment was seen as an integral component in the feedback and advice systems. Chapter 8 has described how the self-assessment statements were derived from the CEFR. A similar process took place with regard to the development of the feedback and advice, although other sources were also used (see Huhta *et al.*, 2002).

Research into the usefulness of the feedback and advice

We have already reported some quantitative research into the value of self-assessment in Chapter 8. In this chapter, we will report on more

'qualitative' research which looks in detail at how users react to the self-assessment and feedback in DIALANG, and, what we can learn about its value and potential. What the studies reported here have in common is that they are case studies of the use of DIALANG and, because of the nature of the methods employed, are necessarily based upon small numbers of respondents, whose reactions are examined in considerable detail. It is impossible to do justice to the richness of the results in a short chapter, but we will attempt to provide some flavour of the data. We end the chapter with ideas for further research into the role of feedback and advice in diagnostic testing, in DIALANG as well as more generally.

The methods employed in the four studies to date (Floropoulou 2002a and 2002b; Yang, 2003; Huhta, 2004) involve detailed observation of users as they proceeded through the DIALANG system, as well as interviews with users, and a questionnaire survey of users.

Floropoulou (2002a) observed, video-filmed and interviewed six students as they navigated their way through an early release of DIALANG Version 1, as part of a usability study, investigating the problems users had navigating their way through the system. At the same time she gathered useful information on self-assessment and immediate feedback.

Self-assessment

Users reported difficulty in responding to Yes/No statements about their ability in the target language, as they felt that the matter was more one of a continuum than a dichotomy. Some would have preferred to have more options, for example 'always', 'sometimes', 'often' and 'never' on some form of Likert scale. One respondent said: '*They are so absolute. The answer is neither yes nor no. There should be some gradation, scales*'. Another said '*A statement like "I can keep up a lively conversation between native speakers" could be answered with "always", "sometimes", "often", "never" or even with "excellent", "very good", "quite good", "poor".*' A third said '*I can do all these things, but I can't answer yes or no. If it asked me how well I can do them I could answer*'. Yet another user had difficulty deciding what he could do and what he could not do. He said '*it depends on the context*'.

Clearly, users have a point: it is rarely possible to be so categorical as to say that one can do a certain thing under all possible circumstances. Future versions of DIALANG might well consider the possibility of adjusting the self-assessment statements to take account of such opinions. The problem would be twofold. First, the link between DIALANG and the CEFR would be loosened, because the CEFR Can-Do statements were couched in categorical Yes/No terms and 'validated' and calibrated as such. The second problem would be one of interpretation. If somebody says that they can do something sometimes, does

that mean they are at a given level, or not? Without further research into what it means to be 'at' a given level, and whether at a level one can 'normally', 'always' or only 'sometimes' do something, it would be very difficult to interpret such responses, however reasonable they may be.

Nevertheless, users responded generally positively to the existence and possibility of self-assessment, and found it useful and interesting, as we shall see later.

Immediate feedback

The availability of immediate feedback resulted in polarized opinions. On the one hand, users found it very useful but others thought that it would interfere with, and somehow influence, their responses:

> *'It is very useful. You need to know whether your answers are right or wrong immediately. . . . it's good to know what you got right or wrong.'*

> *Q*: *'Why did you deactivate the feedback button?'*
> *A*: *'I wanted to get to the very end because I know that typically if you get involved to (*sic*) the process of correction it may influence the way you think, it may change the way you think about the test.'*

> *'I didn't want immediate feedback because it would stop the process and would influence my answers. When you get your answers right you get encouraged, when you get them wrong you get discouraged.'*

These polarized opinions have been observed informally and reported anecdotally whenever DIALANG has been used in workshops or courses. The important thing to note is that users have the option of viewing feedback on each item as they respond, or switching the feedback off and only looking at their results once the test is over. That optionality would appear to be fully justified, but it would be worth researching further whether certain sorts of users prefer to have immediate feedback. Might lower-level learners, for example, find it less useful than higher-level learners? Do users who overestimate their ability, at least in DIALANG terms, tend to be the ones who do not wish immediate feedback? Might a willingness to receive feedback immediately be associated with certain personality types, certain cultural backgrounds, or simply be an individual matter unrelated to proficiency or to social, personality or other variables?

Floropoulou (2002b) investigated learners' attitudes to the self-assessment component of DIALANG, with the hypothesis that there might be cultural differences between her two groups of learners (five Greek and five Chinese speakers). She first interviewed students for their experience of and views about self-assessment. She then observed

them taking a DIALANG reading test, complete with self-assessment, and interviewed them again about the experience afterwards.

She concluded that the Chinese informants were more positively disposed to self-assessment and DIALANG than the Greeks, but there were clear individual differences.Three Chinese students had positive attitudes to self-assessment; although they recognized that it was difficult to do, they believed that it would help learners to improve. However, two Chinese respondents considered self-assessment to be subjective and therefore they would never trust their own judgements.

Similarly, three Greeks considered self-assessment to be a worthwhile activity in order to identify their strengths and weaknesses, while two others had negative attitudes to self-assessment. They expressed considerable scepticism that learners could self-assess accurately.

The Chinese seemed to be less inclined to challenge the value of self-assessment or the tests, whereas the Greeks were somewhat more ready to criticize the tests, and indeed to dismiss the test results if they did not agree with their self-assessment.

The respondents had neither been trained to self-assess nor did they have much experience with it. The most common way they reported doing self-assessment prior to this study was by taking a test and then comparing their answers with the correct answers in the answer key. Such a dependence on external tools, i.e. tests, assignments, teachers' feedback, in order to get a view of their language proficiency suggested that they did not feel secure enough to self-assess. '*We usually think it not very important how you think of yourself*', said one Chinese respondent, who felt that this was associated with Chinese culture. Another Chinese respondent considered that his opinion of his language ability was worthless. Two Greeks claimed that students had to be really mature in order to accurately self-assess and they needed a teacher or a native speaker to correct their mistakes.

One student believed that, while self-assessment was important, it was '*something internal*' and one needs '*external tools to help yourself*'. If learners want to improve, he thought, then developing 'standards' is important. If they want to see whether their standards are adequate, then they should compare themselves with other language learners, in which case taking a test is essential.

Another informant seemed to be very confused about the notion of self-assessment. For him self-evaluation and external assessment were equivalent terms. Although he felt unable to evaluate his own proficiency, his answers revealed that he did indeed self-assess, but was not aware of this.

One Greek respondent considered self-assessment to be the process that followed external assessment. Learners get their results and think of reasons why they have passed or failed an exam. Self-assessment is useful, he thought, but '*learners may have an ideal view of their own ability, so other people may assess them more objectively*'.

Despite their lack of awareness of what self-assessment is, and doubts about its value, many respondents considered the Advisory Feedback to be very useful because it enabled them to understand the levels of the CEFR. They also considered useful the advice on how to improve their skills further. It appears to have been the Advisory Feedback in particular that encouraged some subjects to think about their language ability and therefore contributed to their awareness:

Q: 'So, you disagree with the computer . . .'
A: 'Yes, I disagree with the computer.'
Q: 'A few minutes ago you said that you agree with the computer'
A: 'Yes, and you know why? Because a computer is a scientific device. And now that I know what B1 really means I don't agree. (Reads the description of B2 in the Advisory Feedback.) So, I think my language proficiency is somewhere between B1 and B2.'

When asked to rank the various components of DIALANG in terms of usefulness, Advisory Feedback was in first place, followed by Immediate Feedback. There were particularly negative attitudes to the VSPT (which of course is not part of self-assessment and feedback), while the Explanatory Feedback and the CEFR descriptions of levels were not thought to be too helpful. Interestingly, the self-assessment questionnaire was not seen as a particularly helpful part of DIALANG. The responses of the Chinese and the Greeks were similar in this respect.

Nevertheless, several informants reported having changed the way they viewed their language ability as a result of the DIALANG experience. They believed that they had identified their strengths and weaknesses as a result of the self-assessment and subsequent feedback. In the light of discrepancies between self-assessment and test results, Floropoulou argues, some learners came to realize that their opinion of their proficiency may have been inaccurate.

This study showed that students had not thought about self-assessment in any depth and, therefore, they only had a rough idea of what it was. DIALANG appears to have influenced some at least to consider more seriously the value of self-assessment and a comparison of self-assessment with test results, partly as a result of their experience of the system, but partly as a result of the interviews with the researcher which made them more conscious of the potential value of self-assessment. It is clear that many learners will need to gradually familiarize themselves with the philosophy as well as with actual working methods and the practical use of supporting materials.

Yang (2003) interviewed overseas postgraduate students at Lancaster University immediately after they had taken a DIALANG test, this time with specific reference to the feedback provided, and then again two weeks later, to see to what extent students had indeed followed the advice from the test. Her two main research questions were:

- How do test-takers use the different types of feedback in DIALANG?

and

- What are the factors influencing their use of the DIALANG feedback?

Immediate feedback

Four students received immediate feedback throughout the test, because they wanted to know how they were doing, whereas two who used immediate feedback initially turned it off after having got the first four items wrong – they found this discouraging and it distracted their attention. One student only turned it on for those items whose answers she was unsure of. Six other students preferred to leave immediate feedback turned off, either because they were confident of their responses or because they wanted to complete the test quickly and review their responses later. They also thought that knowing that their answers were wrong would demotivate them. Interestingly, one student said that, since she could not change the answer if it was wrong, she saw no point in using immediate feedback.

Item review

All the students considered being able to review the items very useful, and the classification of items into subskills was also seen as very informative. Two students said that they learned from this what inferencing was, and that specific subskills could be practised. They also felt that they could learn from their mistakes by comparing their response with the correct answer.

Vocabulary Size Placement Test

In general, students found both the score and the short description useful, the score because it 'indicated their position', whilst the description explained 'how rich their vocabulary was'. However, most students felt that one test was insufficient to conclude that their vocabulary was only 'limited' or 'adequate', and several disagreed with the level they had been assigned (believing that it had been overestimated). One student was critical of the fact that the words were presented out of context, and thus the test could not assess how well users could derive the meaning of words from context.

Test results

All but one felt that the feedback on the test result was more informative than a bare numerical score would have been. However, many wanted

to receive a score as well as a CEFR level, especially as, in the words of one student, there was likely to be a large interval between the top and bottom of any band, and she wished to know her position within the band.

This reflects to some extent criticism of the test result in Floropoulou, 2002a. Not all her learners felt that they could interpret the final score. The CEFR description they received was 'not enough to make them understand'. One of her students in particular, on re-reading her result, said that she could not interpret it and she would prefer a score, with an indication of the pass mark or a percentage score instead.

Self-assessment and explanatory feedback

Three of Yang's students barely read the explanatory feedback, because their test results matched their self-assessment. Those students for whom a mismatch was identified did, however, read the feedback in more detail. Five, judged to have overestimated their ability, disagreed with that judgement, since they felt there were many variables which might explain the mismatch (which is, of course, precisely what the explanatory feedback says, suggesting that they had not internalized that feedback). Nevertheless, all but one explored the explanatory feedback looking for possible reasons for the mismatch. One student could not understand why a test should provide learners with such information, especially if the score was based on only 30 items and some of the feedback related to personality issues. The intended tentativeness of the feedback was clearly not apparent to this student.

Interestingly, those students who were told they had underestimated their ability criticized the self-assessment statements for being 'too absolute' (see also the earlier criticisms of the lack of Likert scales). At least one student said that she felt being modest was in any case a virtue. Nevertheless the majority of students said that they had learned from the explanatory feedback, as they had not realized that there were so many factors affecting language learning and their self-assessment.

Advisory feedback

Eleven out of the 12 students interviewed considered the Advisory Feedback to be the most helpful, particularly because of the comments made about the conditions and limitations at each level. All found the advice for improvement useful, although one or two wondered if it might not be more useful for lower-level learners. Some felt they would have liked more specific advice, especially on how to improve their subskills.

Use of feedback

When interviewed later about how, if at all, they had made use of the feedback, one learner, who had been assessed at A2, had made the

effort to read texts outside his academic area. Unfortunately nobody else reported having had the time to follow the advice (which the researcher had copied for them, to encourage them to use it if possible). Half, however, said they would have followed the advice had they had more time (they were busy writing their Masters theses at the time). They all claimed they would have followed the advice if their primary aim in being in the UK had been to improve their English (the respondents were all postgraduate students of content areas, not language students).

In discussing the results, Yang points to feedback models that propose that learners will make more effort to process feedback if there is a discrepancy between their actual performance and their desired goals or their self-perception. This certainly appears to have been the case in the processing of the Explanatory Feedback, and also may help to explain the fact that lower-level learners were more likely to process and follow the Advisory Feedback than higher-level learners, who could be said to have already met their goals.

It was interesting to note that, despite what the DIALANG feedback contains about differences among different tests and the need to consider test results as only one source of information about one's ability, most of the students compared DIALANG and their performance on the test with other tests – particularly proficiency tests and their performance on such tests (TOEFL, IELTS, etc.). They seemed to find it difficult to understand the function of a diagnostic test. Yang points out that theory predicts (Chinn and Brewer, 1993) that learners will filter their reactions to feedback through their beliefs, and in this case, students' beliefs about what a test should be and what it should indicate do seem to have had an influence on how they reacted to the feedback, as well as to the self-assessment component in general.

Theory also predicts (Keller, 1987) that learners will be demotivated if the information given by the feedback does not conform to their expectations, and this may account for some learners' dissatisfaction with Item Level feedback because it did not explain why they got an item wrong. Feedback which convinces learners they can improve if they persevere is more likely to enhance motivation (Andrews and Debus, 1978), which may account for the perception that the Advisory Feedback was considered to be the most useful component in DIALANG, especially perhaps as the differences between levels are juxtaposed in such a way as to make progress to the next level seem feasible.

Most students had believed that the most useful form of feedback would be error identification and correction, which may account for the perception that Item Review and the section on Conditions and Limitations in the Advisory Feedback are the most useful forms of feedback. Some students did not consider the Explanatory Feedback as feedback, nor did they see how it could help them to improve. Since the DIALANG philosophy is that Explanatory Feedback should help learners to reflect on their language learning and set realistic and

appropriate goals, it is important that learners be helped to understand the value of such feedback, and it is likely that mere exposure to feedback alone will not achieve this goal without further training or intervention by teachers or advisers, particularly for those who are not already receptive to the advice or explanations. This will also involve learners coming to understand what the function and purpose of tests like DIALANG might be, since these are likely to be quite far from their ordinary experience of tests – and of other forms of assessment, including peer and self-assessment.

Yang concludes:

> '*The study indicated that elaborated feedback, more than a score, could facilitate learning by identifying learners' weaknesses, helping them realize the cognitive, affective and social factors involved in learning and providing suggestions for improvement.*'

Nevertheless, and recognizing the limitations of her own study, she considers that further research would be highly beneficial, in particular into:

- how effective immediate feedback is;
- how DIALANG users employ the feedback over a longer time span, having them do the test again and comparing their performances to see how effective the feedback is;
- how feedback mediates self-regulated learning;
- how to help DIALANG users to make more effective use of the feedback and maximize the effects on learning of feedback from tests.

Already, further research is in progress, which will hopefully throw light on some of these issues. In particular, Huhta (2004 and in progress) intends to take up some of these challenges. In pilot studies he and co-authors (Huhta and Figueras, 2001) have reported how users react to the CEFR-related feedback in DIALANG in a variety of contexts: Finland, Germany and Spain, in university language centres and language departments, polytechnic universities, centres of adult education, upper secondary schools and private enterprises. He has interviewed teachers and programme coordinators as well as test-takers, and is also conducting a larger-scale questionnaire survey.

Results to date show that users have, in general, positive reactions to the feedback: '*(The feedback) felt quite nice because it doesn't criticize you, it's feedback*'.

The verbal description of CEFR levels was considered more informative than a mere score or CEFR level: '*This is much nicer, that it says (something), than that there is a cold number there*'. However, some comments seem to concern the complex nature of proficiency: a learner comparing levels B1 and B2 said '*They are after all quite near, it can depend on the day*'.

The Test–SA mismatch Exploratory Feedback was felt to be clear and understandable and was interesting for most users, and all said that they would read it if taking the tests for real. '*These are presented in such a friendly way . . . you want to hear the reasons* (for mismatch), *it doesn't say for example that "you don't have confidence in yourself" or "you are too confident", (but) here are these possible reasons so you can sort of search for an explanation that fits you . . .*'

However, in the research setting, the extent to which users read the feedback varied greatly, from one page, i.e. from just one reason, to almost everything. Nevertheless, all users reportedly found at least one reason that would have explained a possible mismatch.

The extended descriptions (in Advisory Feedback) were thought to be clear and understandable, and users felt that they told them more about proficiency than the shorter descriptions (in the Test Results screen). One learner even felt that she could assess her level more easily with the Advisory Feedback than with the short, general descriptions.

In addition, the advice in the Advisory Feedback was perceived as clear and understandable. It was also interesting and useful, although with some reservations: '*. . . fine distinctions of meaning between similar expressions . . . I always play by ear, it's very difficult to find (such information) in dictionaries, dictionaries present that in such a simple way, and then guidebooks on style and writing . . . (it's) very difficult, if you come across something, to find (that) in some guide on style or writing*'.

Some users felt that they were already familiar with the advice, but others found it novel; indeed some users said they were already doing the things advised, but everybody found something new. Interestingly, most users only looked at the advice for their own level and did not explore advice for other levels.

One large question remains, however: would they do as advised?

In the second phase of his research, Huhta plans to explore this issue, with questions like:

- What kind of learners find DIALANG feedback useful? Under what conditions?
- How do the users' views of (typical) **feedback** affect their use of and reactions to DIALANG feedback?
- How do the users' views of (typical) language **tests** affect their use of and reactions to DIALANG feedback?
- How is the 'accuracy' of users' **self-assessment** or difficulty of self-assessment related to how feedback is used?

Clearly, reactions to the feedback and advice are varied, but on the whole reassuringly positive. Nevertheless, numerous research questions remain to be asked, and it is one of the attractions of DIALANG that, since it is freely available on the Internet, such questions can indeed be asked, and hopefully answered.

Moreover, it is to be hoped that explorations of the reasons for the variable reactions to some aspects of the feedback and, in particular, to self-assessment may lead not only to greater understanding but also to the development of activities or even courses that will help learners appreciate the value of self-assessment more and help them to become more self-aware, and thus hopefully lead more learners to benefit from the sort of diagnosis and feedback that can be provided by systems like DIALANG.

Chapter 15: Experimental Items

As mentioned in Chapter 4, DIALANG initially developed items in a fairly traditional format, namely two versions of multiple-choice and two versions of short-answer questions. This was in part because these were by far the easiest to implement in the time scale the Project was working towards. They were also arguably the most familiar to the various item writing teams. However, the Project was very well aware that these test methods were fairly conservative and, indeed, in current language testing practice as well as theory, are regarded with some suspicion, because it is felt by many that they do not necessarily reflect the sorts of things that people have to do with language, that they therefore provide a potentially distorted or simply false picture of a person's ability, and that they may also involve test-wiseness or other sources of construct-irrelevant variance. And indeed a number of item writers, especially in the German ADT, had actually produced some very interesting items using different test methods: they had been deliberately experimental.

Although it was not possible to use such items in Version 1 of DIALANG, it was felt important to show that computer-based testing need not confine itself to using traditional and somewhat boring test methods, that it was possible to experiment with other item types as well and that future versions of DIALANG could, indeed should, implement a wider range of test methods than Version 1 could possibly do. In fact, Alderson had already in 1986 warned of the danger that the use of the computer risked the resurgence of objective item types such as multiple-choice, or cloze tests, and he urged experimentation with novel ways of testing language by computer. In Alderson (1988) a range of different examples were developed to explore what might be possible even within the limitations of computer-scored tests. As mentioned in Chapter 4, these examples were initially developed at Lancaster University for the BBC-B computer in a project funded by the British Council. They were then adapted for use on the IBM PC in a further project which developed LUCAS – the Lancaster University Computer-based Assessment System.

Another reason for developing innovative items was that it became clear early in the Project that it would not be possible to develop computer-scored speaking tests, and so experimentation with alternative ways of assessing speaking within DIALANG was also felt to be desirable.

As a consequence of the desire to show the potential for innovation in test methods within DIALANG, towards the end of Phase 1 of the Project in 1999 it was decided to develop what were initially called Innovative Items, but were then eventually labelled 'Experimental Items'. The Specifications Development Group consisted of Steve Fligelstone (Lancaster University, UK), Graeme Hughes (Lancaster University), Ari Huhta (Jyväskylä University, Finland), Sari Luoma (Jyväskylä University) and Peppi Taalas (Jyväskylä University). The HTML and JavaScript implementation was done by Tom Clapham (Lancaster University).

The Project could not devote much time or resource to this subproject, and so a series of mock-up items were developed using HTML, and produced on a CD for the purposes of demonstration. These items are now available on the DIALANG website at www.dialang.org (click on 'About DIALANG', then 'New item types', or go direct to http://www.lancs.ac.uk/fss/projects/linguistics/experimental/start.htm) and readers are encouraged to visit the website to see how these items work. In practice, whenever these Experimental Items have been demonstrated, both in the 2001 European Year of Languages and at seminars and workshops around the world, they have aroused considerable interest and enthusiasm, and some of the ideas have already been developed into experimental assessment procedures (see Chapter 16 in this volume; Luoma and Tarnanen, 2003; Alderson *et al.*, 2004).

In this chapter, we will examine these Experimental Items, and explore how they might be used to enhance tests and to enable better or more interesting diagnoses and feedback. Many of the items could be incorporated into the DIALANG system in the future; they thus illustrate possible ways forward for the system. By widening the range of item types/test methods available for computer-based diagnostic tests, they also offer possibilities for diagnosis from different perspectives, thereby taking account of the need, pointed out in Chapter 2, for diagnosis to proceed using multiple methods, rather than relying upon one source or type of information alone. In the commentary that accompanies the description of each item type, the possibility of enhanced diagnosis is considered.

The items illustrate:

1 ways in which IT can enhance the quality and authenticity of language test items;
2 additional or alternative types of feedback to the clients (especially at item level);
3 innovative ways of testing direct speaking and writing, i.e. skills that the current system lacks.

There are a total of 18 different 'item types' on the CD/website, namely

Type 1: Pictorial Multiple Choice with Sound
Type 2: Interactive Image with Sound
Type 3: Video Clips in Listening
Type 4: Drag and Drop Activity
Type 5: Reorganization
Type 6: Highlighting/Underlining
Type 7: Insertion
Type 8: Deletion
Type 9: Thematic Grouping
Type 10: Multiple Response
Type 11: Indirect Speaking with Audio Clips as Alternatives
Type 12: Benchmarking in Direct Writing
Type 13: Multiple Benchmarks in Speaking
Type 14: Drag and Drop Matching
Type 15: Transformation
Type 16: Combined Skills
Type 17: Mapping and Flow Charting
Type 18: Confidence in Response

Since these are presented on the CD/website in no particular order, this discussion will look at Types 11, 12 and 13 last and will begin with somewhat more traditional item types. The descriptions are adaptations of the original text on the CD.

Type 1: Pictorial Multiple Choice with Sound

This item type replaces the traditional verbal multiple-choice options with pictures. Users listen to a description (in Finnish) of a young boy and they have to click on the correct picture. When they click on a picture, they receive brief feedback as to the correctness or otherwise of the response. Importantly, feedback to a wrong response (in Finnish, the target language) is in the form of a statement as to why the option chosen is incorrect, not merely that one has chosen the wrong answer. (Clearly the clue could also be in the DIALANG Administration Language System, ALS.) This illustrates the fact that by providing immediate feedback to users, it is possible to incorporate clues as to the nature of the correct answer.

This immediately raises the issue, discussed in Alderson (1988) and Alderson and Windeatt (1991), as to whether it is acceptable for tests to contain clues. In normal tests, giving clues is considered to be cheating, but this need not be the case, especially in diagnostic tests which, like DIALANG, are very low-stakes, or indeed no stakes at all (you cannot fail the test, the results are reported to nobody, and if you cheat you only cheat yourself). Is providing clues as to the correct answer a good thing to do? One might want to argue that if it helps a learner to

learn something, then it is at least useful. The notion that one can learn from tests is not very unusual but it is not often implemented deliberately. However, when feedback (and clues) can be immediately available after a response, then the distinction between a test and a learning exercise begins to become vague, uncertain and somewhat academic.

A further issue raised by the provision of clues (as well as other forms of help, such as online dictionaries, see Item Type 6) and allowing second or more attempts is whether scores should be adjusted for the use of such help. Giving partial credit in normal paper-and-pencil tests is notoriously difficult and error-prone, but the computer has no difficulty in implementing accurately any adjustment to scores that the test designer can imagine (half marks for a second attempt, 1 per cent deducted from the total score for every use of a dictionary, and so on). Whether this is desirable is an entirely different matter, but it is certainly feasible. And recording and then reporting the use of clues and other forms of help could conceivably enhance the diagnosis of a learner's performance.

One advantage of this item type is that the sound, text and pictures all appear together in the same medium when the item is presented on the computer; in the paper version, the sound has to be presented with the help of, for example, a tape recorder.

Possible variations include allowing the learner to listen to the sound more than once, and the immediate feedback could be given in recorded rather than written form. Instead of automatically providing immediate feedback, the item could include a 'feedback' or 'hint' button, allowing the learner to decide whether they want feedback or not. The item could also track how many attempts the learner has made before answering the item correctly and report that to the learner (see Item Type 9, Thematic Grouping, for an implementation of this feature).

The use of pictures and other graphics might enhance the face validity and authenticity of DIALANG items. The item-level feedback would strengthen the use of DIALANG as a learning tool for language learners.

Type 2: Interactive Image with Sound

First users read the instructions and listen to the dialogue which they can hear only once. Then they choose, by clicking with the mouse on the picture on the screen, the room in which they think Herr Müller's friend is. The room chosen becomes highlighted. Users can change their mind and click on another room as many times as they wish. Having made up their mind, they have to click on the Done button, after which they are told if they had made the right choice. Interestingly, there are two possible correct answers to this item.

As with Item Type 1, in the IT environment the audio presentation can be fully integrated into the item without a need to change between two modes of presentation (paper and tape player). The number of

times that the audio file can be heard can also be more easily controlled when presented via the computer.

Another major advantage that IT has over paper-based modes of presentation has to do with the way the learners indicate their choice. They can more easily change their answer simply by clicking on another part of the picture; there is no need to spend time on erasing a previous attempt at an answer since a new choice automatically deletes any visible trace of the first attempt.

Another possible way to interact with a map like this on the screen is to draw lines and routes on the map with the mouse. Rather than just clicking on the place where something can be found, users could be asked to draw the route which is heard (or read). The feedback could then include the particular route described in the text. However, assessment of the route itself, rather than simply the final location, would change the underlying structure of the item and potentially lead to great complexity, particularly in the construction of the task. Such a method might also involve construct-irrelevant variance like users' ability to use the mouse to complete the task.

An alternative approach to providing more feedback could be tied to the type of incorrect answer. For example, if the user had clicked on a room that was in a totally wrong part of the floor, the feedback could remind him/her of the point in the instructions which the user had apparently missed (e.g. 'You were asked to turn right after the lift, not to the left.'), in the hope that this would help users to locate the correct place. Such feedback could lead to learning, or to a diagnostic discussion with a teacher or peer as to why a mistake had been made at that point.

Items such as this could enhance the assessment system in at least two ways. The task simulated by this item, and the skills required in it, are quite common in the real world. Thus, this item type might increase the authenticity and face validity of the assessment system. Secondly, providing an answer by simply clicking on a spot in a picture adds to the flexibility and variety of ways of interacting with test items, which conceivably makes the task of taking a language test more interesting.

Type 3: Video Clips in Listening

Users are first instructed to read the three questions attached to this task, which they can do by clicking on the questions one by one. Then they view the video clip by clicking on the Play button, and they can also pause the clip at any time by clicking on a separate Pause button. They can view the video only once (but obviously more viewing could be allowed). After submitting each answer, users are shown whether their answers are correct or not.

This item could be implemented on paper with the help of a separate video player, but the computer can integrate both in a way that is simply not possible without it. Again, the possibility of controlling the number of playback times is simpler in the IT mode.

The fact that IT makes it easier to have moving pictures as the source material, rather than audio files and/or still pictures, increases the authenticity and naturalness of testing listening comprehension. A great deal of listening in the real world takes place in interaction with people whom we see; the visual context affects our understanding and cannot be fully simulated with audio recordings. Furthermore, visual context may diminish the need for verbal explanation of the context in which the listening is supposed to take place. Thus a more realistic diagnosis of one's listening ability might be achieved by such a multimedia presentation. A possible extension could involve the use of video clips as options in a multiple-choice task, where users are instructed to choose the clip that most appropriately follows on from the conversation.

One variation of feedback could attempt to make a test task such as this a learning task. Before giving the learners the right answer to a question they got wrong, they could be required to go through the task again in the hope that they might come up with the right answer by themselves, although with increasing amounts of help. This help might be made available in the following way.

First, users could be offered the chance of viewing again either the whole clip, or the section which gives the answer to the question they answered incorrectly. If they still cannot answer the question, a written transcript of the conversation on the clip could be made available, possibly with certain key points highlighted in it. In the next step, access to a dictionary could be provided for the learner, and finally, the correct answer(s) could be disclosed, or only the key part of the recording needed to answer the question could be played.

Type 4: Drag and Drop Activity

Users first listen to instructions telling them how to arrange a set of objects on a table or base that the learners can see on the screen. The objects are of different shape and colour (see screenshot in Figure 15.1).

The learners arrange the objects by dragging them, one by one, with their mouse and dropping them into their place on top of each other. After clicking on the Done button, they are told which of the three objects were correctly placed and which were not. If the first attempt was unsuccessful, the learners are allowed to try again until the objects are arranged correctly.

This item type cannot be easily done on paper in a realistic way unless the learners use colouring pens or pieces of coloured paper made for the purpose. IT also allows the provision of immediate feedback and multiple attempts.

This technique can be used for a variety of tasks in which the learners have to understand, on the basis of either a spoken or written instruction, how to put together or arrange various gadgets, machines, vehicles, buildings, maps, objects, pieces, etc. This activity can also be used to rearrange pieces of text (see Item Type 5) or even audio or video clips.

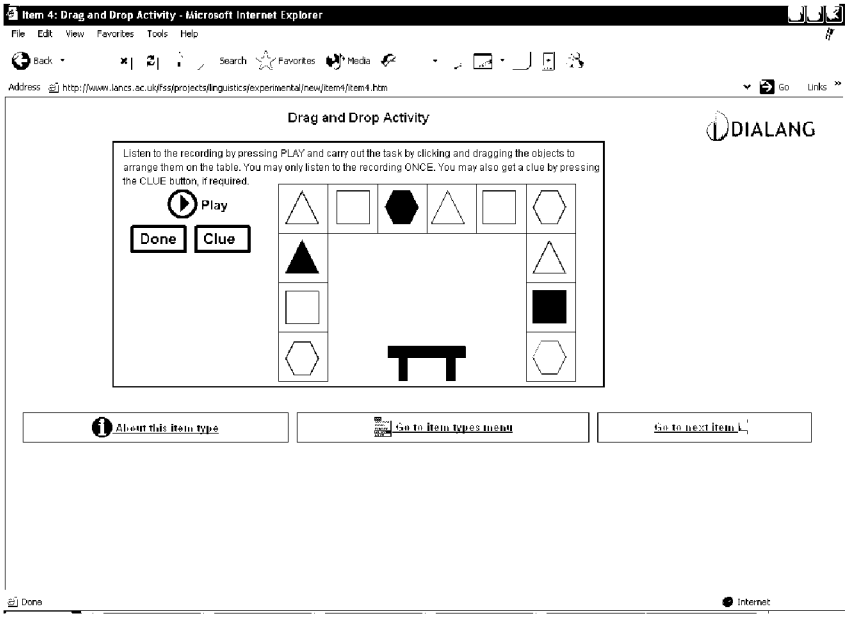

Figure 15.1 Drag and Drop Activity

If users are not able to complete the task after a specified number of attempts, they could be offered a chance to listen to the audio clip again. Feedback can also include the number of attempts which were needed to arrive at the correct solution and the number of times the user listened to the audio clip. The user could also be shown visually how part of the construction is to be completed or where there are missing pieces in the construction. Diagnosis of the problems encountered in completing a complex task could be enhanced by monitoring each step in the process, and by examining where in the process the learner had made a mistake, leading to reflection on why things had gone wrong at that particular point.

This item type could increase the face validity of the system since it would enable many attractive-looking items. Some of the tasks can also be quite authentic as one sometimes has to demonstrate one's understanding of spoken or written instructions by constructing an object or by drawing something.

Type 5: Reorganization

Users have to reorganize scrambled sentences into a comprehensible and coherent text. They do this by dragging the sentences one by one with the mouse and dropping them into their new place (see screenshot in Figure 15.2).

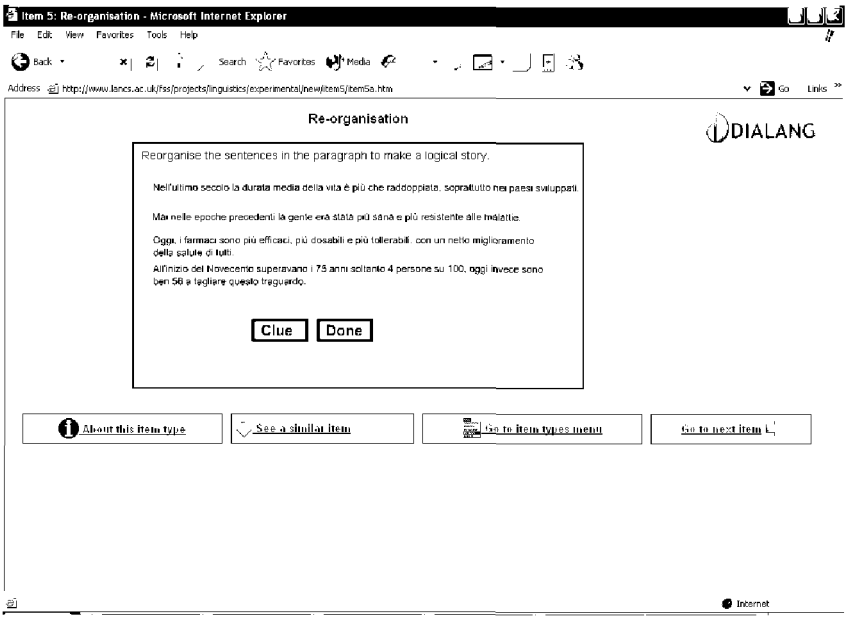

Figure 15.2 Reorganization

Users have access to advice on how coherent texts are typically organized which may help them to see points in the sentences that help them to figure out their original order. After completing the task, they click on the Done button and get immediate feedback on whether the order they put the sentences into was correct or not.

IT allows users to actually move the pieces of text into a new order and see and read the pieces in their new order, i.e. as the new 'text' that they form. This is not possible on paper where users have to try to imagine how the text would read in any other order than the one they see on paper. It might be possible to use a text which is actually cut into pieces which users then try to put together, but even this is not as easy and straightforward as in the IT implementation of this item type.

Reorganization can be applied to other units besides sentences. Words or whole paragraphs can be reorganized to create sentences or texts. This item type would add to the ways in which reading, especially understanding cohesion and coherence, and, indirectly, writing ability could be tested. Users could also be asked to reorganize not simply text, but audio or video clips.

The current feedback only indicates whether the order chosen by the learner is correct or not. Additionally, the user could be told which units (e.g. sentences) are correctly placed and which are not, as is done in Item Types 14 and 17. Demonstration (and diagnosis) of incoherent

organization could be achieved by having the computer highlight cohesive elements that help to determine the correct order of elements in the text, and then seeing if learners could construct the correct order.

Type 6: Highlighting/Underlining

Users have to read a text and decide which sentence contains the main idea. They indicate their choice by clicking on the sentence and on the Done button, after which the program tells them if they chose the correct sentence or not. Users can consult a tailor-made online dictionary which contains the definitions of some of the words or concepts as used in the text, rather than as defined in an ordinary dictionary. The dictionary can be accessed simply by clicking on the Dictionary button on the screen, and then by clicking on the words of interest to display their definitions (see screenshot in Figure 15.3).

The length of underlining/highlighting can be controlled by the test designer: when the learner moves the pointer over a sentence it becomes automatically highlighted (in the sample item the length of the underlined section is one sentence). The learner cannot underline more or less than what is allowed, which removes a major problem for assessing paper-based underlining items.

The learners can change their mind about what to underline more easily since 'erasing' the previous underlining is done automatically by the program.

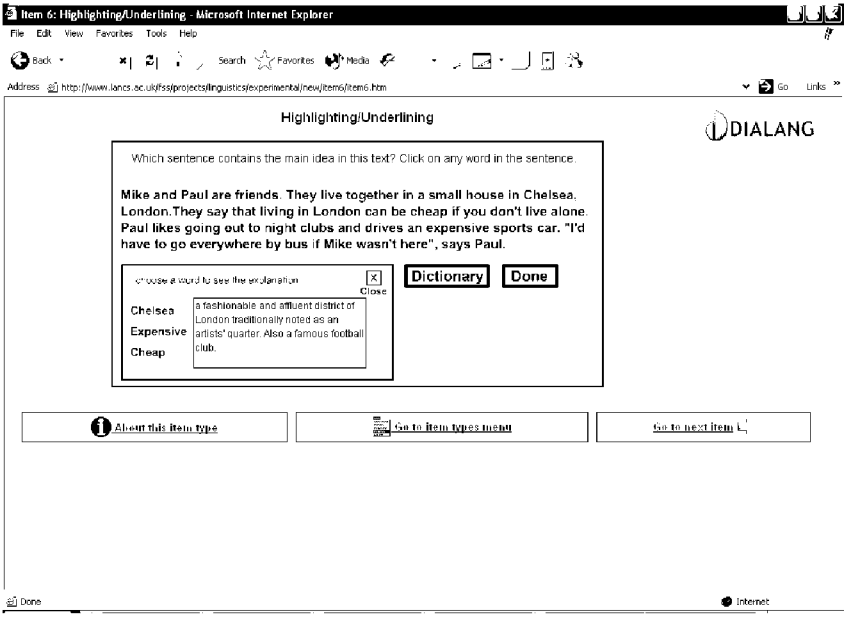

Figure 15.3 Highlighting/Underlining

Underlining/highlighting is a response technique which can be used for several purposes ranging from indicating an answer to a particular question (as in this example) to indicating points where there are errors of different kinds.

The extent and nature of the dictionary help provided for the learner can also be varied, with obvious diagnostic potential. The computer could keep track of the choices made by the learners and whether and how (for which words) the dictionary is used, for example, and this information could be used to develop the diagnostic feedback given by the system.

In addition to merely indicating to the learners whether the section they underlined was correct or not, they could also be guided towards the correct solution, if their answer was wrong, with various kinds of advice. This could include questions which focus the learners' attention on the relevant sections in the text or explicit explanations as to why the chosen section was not correct.

Underlining/highlighting is a technique which could increase the authenticity and face validity of the system, by adding a test method which allows the simulation of an activity which is very common when learners read in order to study and learn from texts.

The use of the dictionary adds to the range and authenticity of items in the system. It also offers the potential to gather information about learners' use of online dictionaries, as well as to adjust scores for the frequency of dictionary use.

Type 7: Insertion

The users' task is to find a place in a (Dutch) sentence where there is a word missing. They click on the gap between the words, and if they have identified the correct spot, a box appears in which they have to write the missing word. They are told immediately whether the word is correct or not.

The IT version of this task enables the task designer to control the completion of this item in a number of ways. The users may be allowed only a certain number of attempts and/or the number of their attempts could be counted and used for research or feedback purposes. A version of the item might have a list of options to choose from rather than a box where the correct word has to be written.

The number of attempts it took to find the place of the missing word could be reported to users. Also, they could be offered access to an online grammar or to a relevant point in such a grammar, to help them to complete the item and to understand the particular grammatical point tested.

This item type would add to the range of different item types that can be used to measure areas such as grammar and might thus increase learners' interest in the system. It also allows the testing of the same

grammatical feature with a variety of test methods, thereby enhancing the reliability of the resulting diagnosis.

Type 8: Deletion

Users have to decide which word (the first sample item) or which sentence (the second sample) does not belong in the text. They indicate their choice by clicking on the word or sentence. If their choice is correct, the extra word/sentence disappears and they are told that they made the right choice. If the choice is incorrect, the program reports this, and allows the learners to try again until they have found the correct word or sentence.

IT implementation allows the learner to actually see what the sentence/text looks like when the extra element has been removed. This is clearly an enhancement in encouraging learners to make acceptability judgements, and in facilitating this process (similarly, the IT implementation of regular gap-fill drop-down items allows learners to see their choice, one by one if they wish, in the context of the text, rather than having to imagine it, as is the case with paper-based multiple-choice gap-fills).

Depending on where and what kind of elements are inserted into words, sentences or texts, these items can focus on different levels of knowledge or meaning, e.g. spelling errors (extra characters in words), grammar (extra grammatical elements in sentences), words or larger units of meaning (extra words, clauses, sentences, paragraphs in texts). The number of elements could also vary; some sentences or texts could have more than one extra element.

It is possible to provide each element in the sentence or text with its own feedback which would explain why it is, or it is not, a good idea to remove that particular element. It is also possible to give the learners questions or other hints to guide their search if the first attempt is unsuccessful.

This item type adds to the range of different ways to test grammar, vocabulary and reading comprehension, possibly also writing (indirectly), which may increase the attractiveness of the system to users and the comprehensiveness of the diagnosis.

Type 9: Thematic Grouping

Users have to decide which words in the list form a group or are directly connected in some way. The number of the words in the group is not given. The users tick the boxes to indicate which words they have chosen, then click on the Done button to submit their answer. The program gives them immediate feedback on their answer. The feedback comes in the form of a statement which appears in a box on the screen and has five different versions depending on the total number of choices and the extent to which they are correct. When the item is correctly

answered, the program also indicates the number of attempts the user made before arriving at the correct answer. It should be noted that (by design) only one of the feedback statements in each of these examples is content-related (and must thus be provided by the item writer). The remaining feedback is automatically generated and is related to how near or far the client is from a correct answer. The feedback in this item thus represents two types of feedback: feedback on the number of choices (how many items are connected) is perhaps more mechanical by focusing on the action that the user has to perform to complete the item (how many such items there are, etc.) and relates to the rubric or instructions of the task. The second type of feedback, one that the item writer has to input, is clearly only about the content/word (or grammar) meanings being tested.

The way in which the elements (words, expressions, sentences, etc.) are related to each other can be varied in a number of ways, depending on the kind of relationships that one attempts to test with the item.

The main benefit of IT implementation is the possibility to provide quick and immediate feedback to users.

The feedback could also incorporate dictionary definitions, even pictures, of the words in the list, which could be provided after a certain number of unsuccessful attempts, as additional support to the user.

Items of this kind may offer interesting possibilities for item level feedback and thus support the learning or diagnosis of the skills or areas of language measured by the particular items.

Type 10: Multiple Response

Users have to choose from a list of words those that form a group according to the instructions. The instructions for the sample item ask users to find the forms of the English verb 'be' among the list of words provided. After making their choice by clicking the boxes of the relevant words, the learners are given feedback which indicates how many of the three correct words they got right (e.g. 1/3 correct). The program monitors the user and if more than three words are chosen, it intervenes to warn the user that there are only three correct words and it asks the learner to de-select the extra words.

This IT implementation interestingly allows online and ongoing monitoring of user responses, even without the user having selected a Done button. This is hardly possible in paper-and-pencil form. It also allows immediate feedback as the learner is responding. Such monitoring is also possible with other item types, and is not an inherent characteristic of this item type alone, but it is of obvious interest for the diagnosis of problems during a relatively complex process, since it may help identify where in that process comprehension broke down or failure occurred.

The possibilities for multiple response items are numerous and are limited only to the extent to which the test designers can find categories

or lists which the learners can choose from on a principled basis. The areas of language tested with this item type can range from very specific and basic, such as the sample item, to much more extensive and elaborate selections (e.g. choosing from a list of summaries or implications the ones that are correct or warranted on the basis of a text).

There are many imperfect implementations of multiple-response items in IT, many of which can be easily 'fooled' by the client simply choosing all the options. The method adopted here, of constraining the number of responses, enables performance on the task to be assessed accurately, allowing partial credit, but without introducing negative marks or other complex scoring models.

This item type is, in fact, an extension of the multiple-choice item type (or a series of true/false statements in a different form); it would add to the range of different ways of measuring a range of skills, which might well make the system more attractive and interesting for users.

Type 14: Drag and Drop Matching

Users have to match words in two different lists with each other according to the instructions provided – in the example provided in the screenshot in Figure 15.4 they have to find words with opposite meanings.

They drag words from the list (or 'container') on one side of the screen, one by one, and drop them beside the relevant words in the

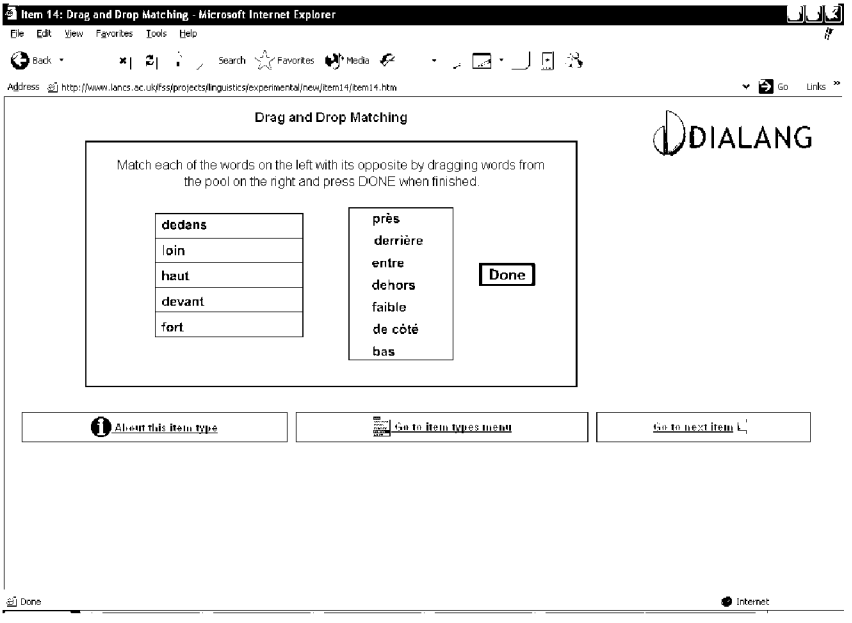

Figure 15.4 Drag and Drop Matching

other list (or 'table') on the other side of the screen. The words in the table cannot be moved, but the words in the 'container' can be moved back and forth as many times as users wish. The 'container' has more words than the table which reduces the chance of getting the item right merely by guessing. After completing the matching, users click on the Done button and are told whether the task was correctly completed or not. If there are any wrong matches, the program returns these to the 'container' – the correctly matched words stay in the table – and prompts the user to try again.

This item type is a familiar one in paper-and-pencil form: the words can be written beside each other to indicate which ones match, or the elements can be numbered and the numbers written into boxes. IT enables users to complete the matching in a way which is familiar to many from modern word-processing programs and computer games. It also enables users to change their minds easily, and indeed prompts learners to do so. Immediate feedback is particularly useful in this item type, since it can indicate which selections were correct and which were not, and may thus lead to self-diagnosis.

Several aspects of the item can be varied depending on how demanding one wants the item to be and which aspects of language are being tested. The technique can be used to test word-level meaning, as in the example provided, or it can be applied to the assessment of broader aspects of comprehension (for example, match statements or explanations and their consequences or implications, based on a text read or heard). Users can be told how the two lists of elements relate to each other, as in the example, or they can be left to figure that out for themselves (see Item Type 9 for an example). The number of distractors – extra words – in one of the lists can also be varied. The computer can keep track of which words users move and where to, as well as the number of attempts, and incorporate this into the feedback.

Users could be allowed access to a dictionary or to hints (e.g. questions, definitions) after a certain number of unsuccessful attempts.

This item type would yet again increase the attractiveness of the system since it would add to the range of ways of testing vocabulary, grammar and comprehension, and it would employ a way of interacting with elements on the screen which is already familiar to many users.

Type 15: Transformation

Users have to transform an expression into an alternative expression that retains the same meaning. In the two sample items, the learners have to write a word in a gap in a sentence, so that the resulting sentence means the same as another sentence that they see on the screen. After this, they click the Done button and are immediately told whether the response was correct or not.

This item type is familiar in paper-and-pencil form. The benefit of IT implementation is that it allows the provision of immediate feedback

and hence diagnosis to users. If the users' first attempts at completing the item result in a failure, they could be provided with help of various kinds, for example the first letter of the required expression might be provided, and if that did not elicit a correct response, the second letter could be provided, and so on. This prompting is not possible in paper-and-pencil formats.

Users could also choose from a list of options rather than providing the answer by writing it (if one aims at testing receptive aspects of vocabulary knowledge rather than both receptive and productive aspects). The technique could also be used for testing other skills, for example in reading comprehension, the learners might be asked to transform the meaning of a text by condensing its meaning into a summarizing sentence or phrase.

This item type can be seen as a variation on the 'text-entry' item type that is already part of the DIALANG system. The inclusion of item-level feedback would be very consistent with the use of the system as a diagnostic and learning tool.

Type 16: Combined Skills

First, users can choose which of the two aspects of reading comprehension, (1) content of the text or (2) words of the text, they want to be tested on. They read a question of their choice, then they read the text, and indicate their answer by clicking on the best option, after which the program tells them if the answer was correct or not. The main benefits of IT are the swiftness of the feedback and the provision of choice and some control over test content on the part of the user.

Possible combinations of two or more skills or areas of language are numerous, and are limited only by the extent to which such skills (e.g. subskills/areas of reading) can be isolated and measured separately from each other.

One type of diagnostic feedback could offer the learners more item-level information, for example about the reason why a particular answer is correct while the others are incorrect. At a more general level, feedback might include information about the role that the skill chosen by the learner might play, for example, in reading comprehension.

Items in which learners can choose the skills tested increase the flexibility of the system from the learners' point of view: not only can they choose if they want to be tested on, for example, reading comprehension, they could also choose which aspects or subskills of reading comprehension they would like to be tested on and get feedback about. This could provide useful follow-up to diagnostic tests like DIALANG where the classification of items into subskills had already indicated areas of difficulty. Users could then choose to explore in more depth their ability on the subskill concerned. Thus if on a reading test feedback at item level indicated problems with inference items, for example, a subtest of this

kind could offer the option of taking more test items that specifically tested the ability to draw inferences.

A very interesting variation on this provision of choice would be to offer the user a choice of responding to easier or more difficult items. This could either be incorporated into the system to replace both the current placement procedures (VSPT and self-assessment) or one of them. In such a case, the user might simply be presented with a menu of levels of difficulty, for example according to the six levels of the Common European Framework, and they would be invited to choose the level of difficulty they felt appropriate. Alternatively, a version of item-level adaptivity could be introduced where, after responding to the first item, with or without immediate feedback, the user could then be asked whether they would like an easier or a harder item next.

Indeed, providing such a set of choices could present an interesting research project, where one could study the effect on choice of providing (or not) immediate feedback. Users' reasons for their choices could also be explored through some form of talk-back procedure.

Type 17: Mapping and Flow Charting

First, users read a text, then they study a flow chart consisting of boxes with a title indicating the kind of content that should go into each box. The boxes represent some of the main ideas expressed in the text and

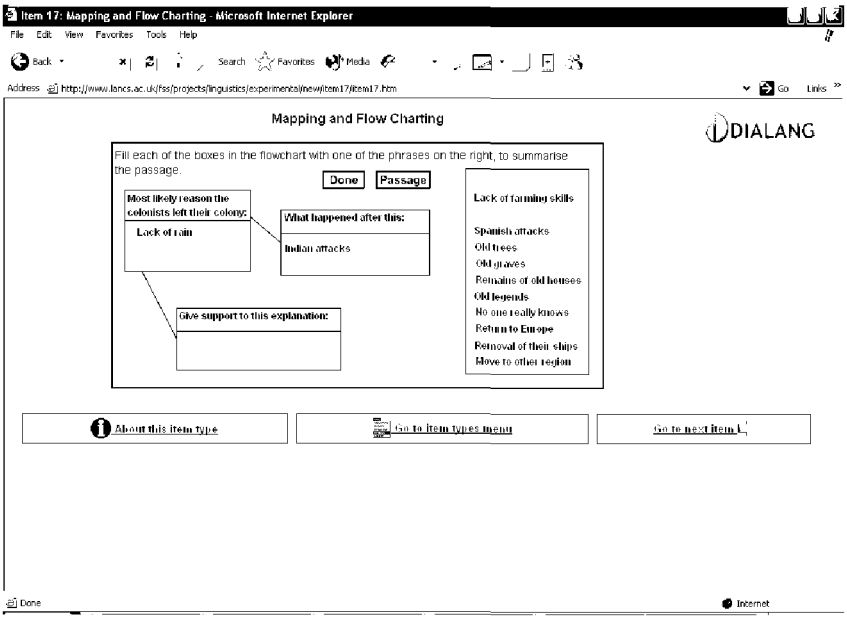

Figure 15.5 Mapping and Flow Charting

their relationship to each other. The users' task is to choose from a list of ideas/propositions listed in a separate box (a 'container') the ones that ought to go into the half-empty boxes of the flow chart. They do this by dragging the ideas/propositions from their container with the mouse and placing them in the appropriate place in the flow chart – see screenshot in Figure 15.5. After this, they click on the Done button and receive feedback as to how many of the boxes they filled in correctly.

This task could be implemented on paper: users can write the ideas into the boxes by pencil, and either erase or cross them out if they want to change their mind. IT implementation of the task simply allows users to do this faster, and to change their mind more easily.

Instead of, or in addition to, filling in the boxes representing ideas, the learners could be asked to give titles or names to the lines or arrows connecting the boxes, which may test more directly their understanding of how the ideas (i.e. boxes) relate to one another.

If the users' choices are partially correct, they could be told which of the boxes were correctly completed; the correct choices could remain in their place while the incorrect ones could automatically return to the 'container' (see Item Type 14, where this is implemented). The learners could also be given advice on how and where in the text to find information which enables them to fill in a particular box (e.g. by highlighting the relevant section of the text).

Items which make use of flow charts, tables, etc. would add to the authenticity and face validity of reading comprehension items since they are often part of texts that language learners face in the real world. They also add to the ways in which main ideas and their interrelationships can be tested: in a flow chart users have to understand and reconstruct several of the main points of the text rather than answer questions.

Type 18: Confidence in Response

First, users respond to a test item, which can be any kind of test question, after which they have to indicate how confident they are about their answer. Confidence is estimated on a simple four-point scale ranging from 'very sure' to 'not sure at all (I'm just guessing)'. The estimation of certainty can cover the task as a whole (as in the first example: the choice of the correct option in a multiple-choice question) or specific aspects of the answer (in the second example, all options in a multiple-choice question are estimated separately). After completing the item and their confidence rating, users receive immediate feedback about the correctness of the response and comments on their confidence. If users chose an incorrect answer and were confident about the correctness of the choice, they are told that they had apparently overestimated their skill or knowledge. If they say they were unsure, and got the wrong answer, they are told that they were right not to be sure. If they get the answer right but said they were unsure about this, they are told that

they 'probably guessed'. If they admit to guessing, they are told they were lucky!

Clearly, the main benefit of the IT implementation of this task type is the immediacy of the feedback.

It is possible to build such confidence testing into many other item types besides the multiple-choice type used in the examples provided. Using user confidence indicators to determine feedback or test scores is not so much an item type as a feature which could be applied to the DIALANG system as a whole. Obtaining confidence ratings could provide more insight into a learner's ability, and might enable affective issues to be taken into account during diagnosis.

Variants on this could ask learners whether they felt the item was relevant to their needs, or too difficult, or unclear, or typical of the sort of problems they face with the language, or whatever. The additional information gathered in this way is probably only limited by the extent to which it might prove useful in diagnosis, either for self-diagnosis or for discussion with a teacher or peer.

As with other items in which self-assessment is required (e.g. Item Types 12 and 13 below), this item requires and may develop self-assessment skills. Some researchers claim that the addition of confidence ratings has the potential to gather more information about users' ability than the same item without this feature. Encouraging users to be explicit about their level of confidence might make them more self-aware. Whether it would encourage or discourage guessing is an open question.

An alternative to marking one's confidence by ticking boxes would be to use a slider which can be moved across a line, relating the position of the slider to a percentage figure: placing the slider in the middle of the line would indicate that the learner is 50 per cent certain that the option is the correct one. Moving it towards the end of the line would either increase or decrease the percentage certainty. In the case of the second example, the other sliders could be linked with each other, so that increasing the value of one would automatically decrease the values of the others (the total remaining constant at 100 per cent).

The productive skills

Testing the productive skills of speaking and writing is at present virtually impossible in a computer-scored system like DIALANG, although there are interesting developments in this area. Educational Testing Service (ETS) is developing a system called e-rater (http://www.etstechnologies.com/) which rates native speaker compositions, and Phone-Pass (http://www.ordinate.com/) uses a computer algorithm to assess a person's speaking ability on a number of tightly constrained item types. Doubtless there will be interesting developments in these areas in the future.

DIALANG has developed methods to test writing indirectly, and these are implemented in the traditional item types in Version 1. One interesting technique to assess speaking ability indirectly is the following item type.

Type 11: Indirect Speaking with Audio Clips as Alternatives
Users read the instructions, then listen to a recorded message (in German) from the Cable TV service. After this they listen to four different messages that they might leave on the service's answering machine and decide which of them is the most appropriate for the purpose. Having indicated their choice by clicking on one of the options and the Done button, they are told whether the message they chose was the best one for the particular purpose.

IT makes it possible to integrate the written and audio media into the same item and thus simplify the testing process. The number of times users can listen to the input message and the options can also be controlled when the item is administered on the computer. And, of course, immediate feedback on the correctness of the choice is available.

Items of this kind would add to the ways in which listening comprehension is tested in DIALANG and they would also enable the system to test speaking in a somewhat more authentic way than totally paper-based indirect items could do.

However, DIALANG has experimented with two other ways of assessing the productive skills, by using benchmarking and self-assessment. In essence, the following two item types involve users comparing their real or imagined performance on a speaking or writing task with a set of written or spoken performances which have been elicited from learners doing the same tasks, and which have been rated by expert judges as being at particular levels on the Common European Framework.

Type 12: Benchmarking in Direct Writing
Users first have to read the instructions and write a text in response to a writing task. Afterwards, they can see examples of texts written by other learners on the same task. Users can control which performances they see and compare them with their own. They are encouraged to judge which of the examples (benchmarked by over 30 judges) most closely resembles the one they have written. They are also provided with a range of comments on the sample answers, which include explanations for why the performances have been rated at the level they are given, and extracts from the Common European Framework which describe what sort of texts a writer at that level can typically produce and what the characteristics of their writing are, as well as under what conditions and limitations such texts are produced. Users can also explore the written performances for the grammatical, vocabulary and spelling

problems/errors they contain, which are marked up in a variety of different ways, and they can view a number of tips which give advice on how best to go about tackling the writing task, in an attempt to raise users' awareness about the demands of the task and to help them evaluate how successful their own performance was.

This IT-based item type allows the presentation of a number of examples and a variety of feedback or advice in flexible ways and instantaneously. To do this on paper would require a whole booklet of dozens of pages, and the navigation within the booklet would be slower and more difficult than on the screen.

Unlike indirect writing items, items such as this allow the learner to actually produce language, making the testing more direct and authentic compared with indirect items. The item type also requires the ability to assess one's own writing, as well as aiming at improving the user's self-assessment skills.

Currently the feedback is related to the benchmarked examples, not the text that the learner produces. The feedback features allow the learner to get access to overall advice on what to pay attention to when writing the text for this particular purpose as well as specific linguistic problems in the benchmarked items. In addition to problems, various positive aspects of the examples could be shown, such as particularly good and idiomatic expressions (words, structures, sentences) for the particular level the sample is a benchmark for.

Initial studies on this item suggest that this kind of comparison may be difficult for beginning learners who may not easily spot particular problems or good points in their own writing, even if they see such problems displayed in the benchmarks. Ideally, this item type should be able to display various kinds of problems (and good points) in the client's own text. This would require an inbuilt spellchecker, grammar and vocabulary analyser in the DIALANG system, which has not been feasible to date but which may be possible, at least partly, some time in the future, thereby greatly enhancing computer-based diagnosis.

This item type also offers the possibility of improving the learners' writing by requiring learners to rewrite their texts according to information or feedback that they get on the benchmarks (or, if possible, on their own first draft text).

DIALANG lacks direct tests of speaking and writing. Items such as this would fill a major practical and conceptual gap in the system. Such items also fit very well into the overall, learner-centred approach adopted in DIALANG, which encourages self-assessment as part of awareness-raising. This item type is based on self-assessment, it can help improve both self-assessment and writing, and it would introduce into the system a feature which it so far lacks. It also has rich feedback and pedagogic possibilities. Chapter 16 explores in more detail the use of this technique in a couple of experimental studies.

Type 13: Multiple Benchmarks in Speaking

Users first read the instructions which ask them to think about how they would respond to a tourist who asks which places they would recommend visiting on a sightseeing tour of the users' local town or village. Then they think about their answer, composing it mentally. When they are ready, they listen to a series of speech samples, presented in random order, in which other learners give their response to the tourist. The users' task is to compare their own possible responses with each of the samples and to say whether they think they can do the task better or not. As in Item Type 12, the samples are benchmarked performances rated by a panel of judges.

On the basis of the users' responses, the program then works out the level that the users are probably at, and reports that to them. There are 20–25 learner responses in total, but the user is only obliged to listen to the first seven before the program can calculate their level, and then allow users to see their result or to continue listening to the benchmarks. Users also have the opportunity to see a verbal description of the speaking ability of learners who are at that particular level, based on the CEFR and characterized in terms of types of speech, characteristics of speaking and the conditions and limitations of speaking at that level.

This IT-based item type allows the flexible and rapid presentation of a number of samples. To do this on paper would also require a tape recorder or other sound device, which would complicate the administration of such items.

Items such as this allow the user to actually produce language, making the testing more direct and authentic compared with indirect items. The item also requires an ability to assess one's own speaking, but hopefully may also lead to improving the learner's self-assessment skills.

Users could be asked to make a recording of their own answer on tape or on the computer they use for taking the DIALANG test, and then to listen to their answer before proceeding to the samples. Hearing one's own response, rather than imagining what one might say, may well make the comparison easier and result in more accurate self-assessment.

It may also be possible to give the learners advice while they are making the comparisons to make the self-assessment easier. This would require, however, more research into the basis of self-assessment in these kinds of tasks.

Conclusion

In this chapter I have hopefully illustrated ways in which the DIALANG diagnostic system, as well as diagnostic systems more generally, could be enhanced by the use of more innovative item types, coupled with the provision of a range of help, clues, feedback and tools to aid reflection, such as self-assessment, confidence rating and the like. Clearly, the development of templates to allow the authoring of such experimental items would

need careful consideration. Each of the item types illustrated here was individually crafted, and in a larger project would require careful design, storyboarding, piloting and experimentation with delivery in various environments before they could be implemented. Whether such item types and tools would indeed enhance diagnosis remains to be seen, but it is surely worth the effort to work towards such enhancement, if diagnostic tests and procedures are indeed to become the interface between assessment and learning. In the next chapter, I will explore some recent research into some of these item types and procedures, as well as examine users' reactions to the feedback they receive from such diagnostic tools.

Chapter 16: Experiments in the use of self-assessment

As discussed in Chapters 8 and 14, self-assessment is a central component of DIALANG, because it is believed that without self-assessment there can be no self-awareness, and self-awareness is important to language learning, both in terms of knowing what level one has already achieved, and in terms of knowing one's strengths and weaknesses, and one's preferred way of learning and developing. The DIALANG philosophy of diagnosis is that a comparison is essential of one's self-assessment in any given aspect of language use with one's performance on a diagnostic test which is not high-stakes, and which therefore risks minimal distortion of one's ability. Such a comparison may reveal interesting discrepancies between one's self-assessment and the test result. Which is not to say that the test is correct and the self-assessment inaccurate, but rather that two sets of information about one's ability are better than one set. That is why in the Feedback and Advice section of DIALANG there is quite an extensive discussion of reasons why self-assessment and the test results may not match – again, to help provide the user with further insights into their language competence.

However, there is another potential use for self-assessment in a system like DIALANG, and we have described this in Chapter 15 when looking at Experimental Items. There we presented a proposed means of assessing productive language skills like speaking and writing, which is currently extremely difficult in DIALANG. Since DIALANG is a computer-scored system, and not merely a computer-delivered system, direct tests of one's ability to speak a foreign language, or to write in that foreign language, cannot currently be implemented in DIALANG. Instead, as seen in Chapter 15, DIALANG has developed two item types (numbers 12 and 13 in that chapter) which use self-assessment as a central component in the test method. In essence, users are invited to produce written or spoken text, and then to compare those texts with texts produced by other learners which have been rated by expert judges in a so-called benchmarking process. Users are invited to

consider whether they can do (or have done) better than the performances they view or hear, and are thereby encouraged to arrive at a self-assessed estimate of their ability in terms of the CEFR.

This item type, which we believe is quite innovative, has aroused a considerable amount of interest whenever the Experimental Items have been demonstrated. Unfortunately, the pressures of completing Version 1 of DIALANG and developing Version 2 were such that it has not proved possible to develop these two item types into tests of writing and speaking in the DIALANG system, although there are hopes that this might be possible one day.

However, others have been able to take the basic idea as exemplified in Item Types 12 and 13, and experiment with them, with interesting results.

Self-rating of speaking ability

The ideas which were developed in Item Type 13, the multiple benchmarks in speaking as presented in Chapter 15, came originally from a Masters thesis by Dandenault (1997), as reported in Luoma and Tarnanen (2003). Dandenault used a self-rating instrument of speaking ability with 149 French-speaking students who were learning English. She presented the learners with eight short samples of learner talk at different levels of ability, and asked them whether they thought they could do better, worse, or about as good as, the samples if they were to do the same task in English. A week later they were also asked to self-assess their overall ability in the language, as well as their reading, listening and writing abilities. The self-assessment scale comprised seven points, from 7 – 'Can express ideas in English like a native speaker' to 1 – 'Cannot express ideas in English at all'.

The task involved telling, in their own words, the story represented in a wordless cartoon storyboard. The internal consistency reliability was .81, which is high for a test of eight items only. There were very few cases where informants rated themselves better than one speaker at a given ability level and higher or lower than another speaker at the same supposed ability level. There seemed to be no gender effects in the data: '*Men and women did not significantly differ in how consistent they were in the way they self-assessed against the various voices on the stimuli tape*' (*op.cit.*, pp. 57–8). Nor did there seem to be any difference in the consistency with which students at different levels of ability self-assessed.

There was a very high correlation (.97) between the ratings of the voice samples by native English-speaking judges and those made by the second language learners. Dandenault suggests that this indicates that the learners were making judgements in a way that was highly consistent with how native English-speaking judges made judgements. Although there was only a .54 correlation between the informants' scores on the self-judgement task and their self-ratings on a single

7-point Likert scale of ability in speaking, Dandenault suggests that this may be due to the substantial differences between the two instruments, since the self-rating '*may elicit responses that are somewhat more variable due to factors like reliance on memory, interpretation of what it means to be able to express oneself, selection of different reference points by different informants*' and so on (*op.cit.*, p. 61). The fact that informants' results were highly similar to the judgements made of the stimulus voices by native-speaker judges gave her reason to believe that the more focused benchmark rating tasks had greater validity as a measure of (self-rated) speaking ability. She concluded that the self-rating task had '*strong face validity and strong internal consistency*' (cited in Luoma and Tarnanen, 2003: 443), which encouraged her to propose further studies with a larger number of samples, as well as the conducting of comparisons between informants' actual ability as determined by their performance on the story-telling task, and their self-rating. It is not known whether such studies were subsequently conducted.

Self-rating of writing ability

Luoma and Tarnanen (2003) report on a study, where the reactions to a self-rating procedure of writing ability were gathered from six learners of Finnish as a second language, using the ideas explored in Item Type 12 in Chapter 15 as the basis for the procedure. The questions they address include the correspondence of learner self-ratings with teachers' ratings, the nature of the self-rating process and the extent to which the program supports the learners' writing processes, what learners focus on during the self-assessment and the value of the self-rating and benchmarking system. The learners' screen actions were video-recorded and the learners themselves were interviewed immediately after the self-rating procedure was complete. They also gathered teacher assessments of the written texts produced by the learners as part of the self-rating procedure.

The two tasks were:

Task 1. *Write a letter to your Finnish friend. Tell about the following:*
- *how you are*
- *how your family is*
- *what the weather is like*
- *what you will do on your holidays, which are just beginning*

Task 2. *Look at the picture. Write a brief story that tells*
- *what has happened before the picture*
- *what is happening in the picture*
- *what could happen next*

Note. Your story can be about whatever you want, as long as you use the picture in your story.

(Luoma and Tarnanen, 2003: 463)

Learners were asked to input their responses to the first task into a box on a computer screen. They were then asked to move to the next screen, where they were to examine each sample response in turn (each having been benchmarked to one of the six levels of the CEFR), inspect the comments about the sample texts provided in the system, and then decide on their own level by comparing their own text with the benchmarked samples. Thereafter, they continued to the second task and repeated the process.

The benchmarked samples had been rated by ten raters, on the six-level CEFR/DIALANG scale, and for each task, six scripts were selected where there was greatest consensus among raters, and which were thought to best represent the particular CEFR level. These scripts were then analysed in order that brief comments could be written about the scripts. The aim was to present comments that illustrated the scale, and that helped learners pay attention to relevant features when rating their own texts. The authors report using three main categories in their comments: comprehensibility, complexity and accuracy. Some comments were grouped under three headings: General; Strong Points; Weak Points. Other comments were tagged directly onto the benchmarked texts. These tagged comments related to four aspects of the texts: comprehensibility, vocabulary, structures and spelling.

Similarly to Dandenault (1997), Luoma and Tarnanen report that their six learners found the self-rating task interesting and easy to do and they found the whole experience positive. Time on task varied greatly, from 18 minutes to more than an hour for the first task, and from 15 to 40 minutes on the second one – all took less time on the second task – and some reported that however enjoyable the experience, two writing tasks was the maximum that should be required. Whilst writing, however, the learners appeared to be quite absorbed in the process and at ease with the use of the technology, possibly because they were not taking a real test but were seated at a computer on their own, much as they might be when writing an email letter to a friend.

Unused to seeing texts written by other learners, they found it instructive to see what a good performance on the tasks looked like. The authors speculate that this may have led learners to look at their own texts differently. They started the assessment process from the lowest levels and then proceeded up the scale until they found texts that they felt were similar to their own. But even those who assessed themselves at intermediate levels also read the higher samples out of curiosity to see how good the texts were. Despite these positive results, learners reported finding it difficult to decide between the two benchmarks closest to their own text, and they all said they would like to know how a trained assessor would assess their own texts. Unfamiliar with searching for mistakes in their own texts, they reported difficulty in noticing their mistakes and in imagining how well a reader would understand their text.

The learners' self-rating corresponded quite closely with the teachers' ratings, with one-third being exact matches, and all but one of the mismatches were less than one band on the six-band scale. Interestingly, most of the mismatches were overestimations by the learners.

When learners assessed the benchmarks, they paid particular attention to content and to language, commenting on salient mistakes, albeit without metalanguage. They also, unsurprisingly, paid attention to the length of the benchmarked texts compared to their own.

Luoma and Tarnanen also analysed the learners' texts in the light of the teachers' and the learners' comments, and in comparison with the comments they had prepared on each benchmark. In the light of these comments and their analyses, they suggest that their initial categories of comprehensibility, complexity and accuracy might be less helpful to learners than other categories, including how learners say what they want to say. They conclude that for such a promising self-rating system to be useful to learners, more work is needed on exploring a commenting structure that is comprehensible to learners and useful for them. '*A challenge in building this sort of tool, however, is that it guides learner thinking. The developers are responsible for the quality of the program's conceptual basis, which is reflected in the tasks, benchmarks and comments that the tool contains*' (2003: 461). Indeed.

Use of self-rating in stylistic analysis

In a field unrelated to foreign language testing – that of the teaching of stylistics to undergraduate native speakers of English – a team of linguists and systems developers have taken the basic idea of Item Type 12, Benchmarking in Direct Writing, and applied it in an undergraduate course at Lancaster University (Alderson *et al.*, 2003).

First-year undergraduate students of English Language and/or Linguistics in the Department of Linguistics and English Language at Lancaster University take an introductory course in stylistics, labelled LING 131. This course involves an understanding of the basic linguistic tools of stylistic analysis and the detailed application of these tools to the analysis of the three main literary text genres – poetry, drama and prose. In order to ensure the development of the skills of stylistic analysis, students need considerable practice in analysing texts as well as feedback on their achievements. In a typical one-trimester university course such as LING 131, there is little time to do intensive stylistic analysis, or to practise the use of the linguistic tools, and even less time to give students feedback on their efforts. Mick Short and associates have, however, developed a Web-based course, 'An Introduction to Language and Style', which covers the same ground as the traditional lecture and workshop course, but using information technology (see Short and Archer, 2003, for details). One of the perceived advantages of such a course is that it enables students to access the input and do the various

exercises at a time and place of their own choosing. They are thus free to practise stylistic analysis at length. Short is currently investigating whether such a Web-based course does indeed offer significant advantages over the traditional mode of delivery. Preliminary results suggest that students see both advantages and disadvantages in the Web-based course.

However, one of the dilemmas of Web-based courses is how best to assess the students' progress, and, all too often, Web-based courses which are supposed to reduce teaching loads result in very heavy loads of assessment of students' work. At the same time as work commenced on LING 131, the software for DIALANG was being developed at Lancaster University, in the same department, and as a result of presentations of some of the DIALANG ideas, it was felt that cooperation between the two projects might be mutually beneficial. Specifically, it was agreed that the application of Item Type 12 to the assessment of students' writing in the Web-based LING 131 course could be of considerable interest and benefit.

The concept of Item Type 12 was thus examined, and a template developed for use in LING 131. DIALANG programmers first mocked up the template, and then the Web-designer for LING 131 significantly enhanced the interface of the item type to match that of the developing Web-course. The basic idea was that, in preparation for their own end-of-course coursework assessment (which takes the form of a stylistic analysis of a poem, or of an extract from prose or poetry), students would be asked to do a stylistic analysis of each of the three literary genres, as part of their formative assessment, and input their analysis into the Item Type 12 system. They would then be able to compare their own work with that of previous Lancaster undergraduates who had already analysed the same texts for a real coursework assessment. Those stylistic analyses would be benchmarked by a team of expert judges and so the students taking the Web-course would be able to compare their own performances against the benchmarked performances and reach a self-assessed estimation of their result, not in terms of the CEFR, of course, but in terms of regular UK undergraduate degree classifications.

The aims of the project were to provide formative assessment for the LING 131 Web-based course, to allow students to develop their analytical and essay writing skills and to give the students the opportunity to compare their work with that of past students. In addition, it was hoped that students would improve their ability to self-assess their work in the light of the feedback and guidance given by the system.

The first step in the development process required students on the traditional stylistics course, in parallel to which the Web-based course was being developed, to submit, as one of their pieces of assessable coursework, an analysis of a particular poem (George Herbert's *Easter Wings*), an extract from a play (*The Important of Being Earnest,* by Oscar Wilde) and an extract from a novel (*Great Expectations*, by Charles

Dickens). Students were requested to analyse these texts in relation to aspects which are important for an adequate analysis of the relevant text, and also to relate their analysis to their own general interpretative remarks – see general impression, below. The aspects of analysis were:

Poetry
- general impression
- graphological deviation
- syntactic parallelism
- vocabulary

Prose
- general impression
- point of view
- speech presentation
- syntactic parallelism

Drama
- general impression
- turn-taking and power relations
- politeness
- Grice's cooperative principle

Students taking LING 131 in 2000/01 were asked to contribute to the development of the self-assessment mechanism by giving the project permission to use their essays (already submitted as part of the regular assessment process, they were thus not produced specially for this experiment and are fairly typical of the work produced on the traditional course). They were given extra feedback on their reports as a reward for allowing the use of their essays. Fifty-three essays were collected in total.

The 53 anonymized essays were then printed out, and independently marked by the two course tutors, and a former tutor on LING 131, using the criteria normally used on this course. At the same time, the markers were asked to make written comments to justify or explain the marks they had given.

The results were collated, and 24 essays were selected (eight from each genre), based upon the amount of agreement among marks, and among the comments made by the markers. These 24 essays represented a range of marks across the marking scale used in the Department, and traditional in other UK universities. They were then sent to colleagues in other Higher Education institutions who were teaching similar introductory stylistics courses, and they were asked to mark the essays in their normal way, and make written comments on the essays. In the event three external and two further internal markers returned their marks and comments.

At this point, short but significant extracts were chosen from the students' essays, based upon the marks given and the written comments, and commentaries were written on these extracts, again based upon the marks given and the markers' comments, and amplified by the course tutors taking part in the experiment. Since the experimenters did not wish to use the full-length essays produced in the traditional summative assessment, but rather to provide students with formative, and arguably diagnostic, feedback, it was felt both more appropriate to ask students on the experimental course to produce shorter pieces of work, and more practical for them to read extracts of essays rather than the full-length essays. Thus the selection of the extracts and the devising of the self-assessment task had to go hand-in-hand with the development of the commentaries that students would see on the extracts of the benchmarked essays.

Once the extracts of the benchmarked essays were ready and commentaries had been drafted, the self-assessment mechanism itself was then designed. Students would be asked to complete three self-assessment exercises, one each on a poem, an extract from a play and an extract from a novel – the same texts as had been used for assessment purposes in 2000/01. Each exercise was broken down into a series of sub-parts.

Students were asked to complete the exercises either during the timetabled workshops, or in their own time (usually in their study bedroom or in departmental computer labs) and then to input their answers into the system.

Students were then encouraged by the program to view the responses of previous students, and to read the commentary on those responses. Students were free to scroll through these commentaries and extracts as they wished.

Since this was an experiment, students were not obliged to do all three self-assessment exercises, but they were asked to complete at least one analysis for self-assessment, and to input it into the system.

The evaluation of this experiment, itself part of a much larger experiment in the use of the Web to deliver a stylistics course, took the form of a series of observations of students using the system, questionnaires answered by all students at the beginning, middle and end of the course, interviews with individual students, written comments by students on a voluntary basis and focus group sessions conducted by an outsider (not the course tutors).

Results

In the event, data specifically relevant to the self-assessment experiment was gathered from 40 of the 79 students who took the course. Of these, 33 said that they had found the self-assessment tasks useful in preparing

to do their coursework assessment. The majority of respondents said they had found it provided good practice, and helped them to see where they needed to improve and to identify their weaknesses. Several appreciated being able to monitor their progress, and to compare their own work with that of other students. Of the few students who said they did not find it useful, several had simply not done the analyses, and one would have prefered his/her work to have been marked by a tutor.

Reports from course tutors confirmed that a number of students did not bother to do their own analyses. Instead they simply scrolled through the comments on the benchmarked analyses. Thus they did use the self-assessment mechanism but not as intended. There are doubtless a number of possible reasons for this. Either the students were simply too lazy or busy with other things to do the analyses – they were not obliged to do the task, after all – or they simply did not see the value of being able to compare their own work with that of others.

Nevertheless, it appears that many students appreciated being able to see other students' work and the marks the benchmarked analyses had been given, as well as the comments written on the analyses. However, most students only completed one out of the three self-assessed analyses, and fewer bothered to input them into the system for systematic comparison with the benchmarked analyses. Students should perhaps not be allowed to decide whether or not to attempt the stylistic analyses, since an important part of the self-assessment task is for students to reflect on what they have themselves produced, and then to compare it with other people's work and the established standards. Only then can they truly assess their own work. It may well be the case that first-year UK undergraduate students, many of whom are well known to be more interested in sex and alcohol than they are in their academic studies, need more encouragement to self-reflect, and indeed may need some form of training or induction into self-assessment and its value, before they are asked to do a task which might be thought to require a degree of maturity, self-awareness or simply motivation.

In a re-run of the experiment in the academic year 2003–2004, all students were obliged to do one self-assessment exercise, the one on poetry, and were asked for their opinion about the value of this at the end of the course. The results were very encouraging. Of the 66 respondents, 63 (95 per cent) said it was useful to be forced to do the poetry self-assessment, and none disagreed. When asked to explain, the following were amongst the responses:

> *It made us revise (12) work (10) think (4) learn (2) understand (2). It was motivating (3). Good preparation for CWA (10) except no feedback (1). Good to be able to compare with others (4).*

Although 44 (67 per cent) only attempted the one exercise they were obliged to do, 16 completed two and four actually did all three. When

asked whether the self-assessment exercises had been useful in preparing to do their coursework assessment, 82 per cent (54 students) responded positively and only 9 per cent said No. Reasons given included the following:

> *Good to have preparation/model for CWA (17). Good practice/help (10). Increases understanding and insight (4). Relates to previous ideas (3). Compare with others (2). Shows you where your weaknesses are (2). Draws answers from us (1).*

These encouraging responses suggest that self-assessment could work well, if ways could be found of ensuring that students make the effort to do the analyses and then input the texts into the system in order to compare their work with the benchmarked analyses and the comments made on those benchmarks.

In this chapter, we have seen how the idea of self-assessment against benchmarked performances, as developed in the Experimental Items presented in Chapter 15, has been implemented in various contexts, and we have discussed how diagnosis might be enhanced through the application of these ideas. The incorporation of these techniques into a course is of particular interest, as it illustrates how diagnosis can indeed form an interface between learning and assessment.

Nevertheless, it is clear from the research reported here, as well as from that reported in Chapter 14, that learners cannot simply be given self-assessment tasks and be expected to know how to do them, how to evaluate the results, and how to derive maximum benefit from the experience. It is clear that many learners are ignorant of or sceptical about the value of self-assessment, and need to be inducted into the philosophy and principles that lie behind the approach, and they need to be guided in their use of self-assessment tools as well as encouraged in a variety of ways to reflect upon their own learning.

In short, self-diagnosis does not come easily to all learners, and whilst it will remain a goal of many educators as well as of projects like DIALANG, it is clear that more attention needs to be paid to convincing learners of the value of such approaches.

Although self-assessment may be an important component of a diagnostic system, such systems need not depend upon self-assessment, and there is clearly room for a variety of different approaches to diagnosis which may still be learner- and learning-centred but which use more traditional approaches to identifying strengths and weaknesses in a learner's performance or knowledge. Test-based approaches are clearly one alternative amongst others, and although, as we have seen in Chapter 2, diagnosis often proceeds in an individualized manner, it need not. Computer-based diagnostic assessment is likely to become increasingly sophisticated in the future, through the use of novel elicitation

techniques, some of which we have explored in this and the previous chapter, as well as through better understanding of the development of language proficiency and the processes that lead to such development. In the final chapter, we will attempt a look into the future of diagnostic testing and asssessment.

Chapter 17: The future of diagnostic testing

In this book I have discussed at length the need for diagnosis of language learners' strengths and weaknesses, I have commented on the remarkable lack of treatment in the language testing literature of how diagnosis might most appropriately be conducted, and I have lamented the absence of truly diagnostic tests of foreign language abilities and the lack of research in general into this area. I then described the DIALANG Project, and explained how it is unique, not only for addressing the need for diagnostic tests, for its complexity in terms of number of languages and tests, and for being based on the Common European Framework, but also for its incorporation of self-assessment into the system, for its provision of extensive feedback to learners and for its innovative delivery of tests over the Internet.

Computer-based testing is particularly suitable for diagnostic testing, because it offers the possibility of immediate feedback. Moreover, delivering tests over the Internet means that individuals can take a test at the time and place of their choosing. Thus individualized diagnosis when the learner wants it is an added benefit of Internet-based testing.

Given these advantages, together with the fact that more tests are being delivered by computer these days, it is quite possible that the future will see more truly diagnostic tests than are currently available. However, for such diagnosis to take place, we do not merely need enhanced technology; we need a much better understanding of the nature of language proficiency and how it develops. Studies of second and foreign language acquisition are still in their infancy, and have not yet contributed much to an understanding of the development of proficiency. The empirical evidence for foreign language development is still fragmentary, though we have detailed knowledge of the development of isolated aspects of the language, and studies of development tend to cover a limited range of the developmental path. Moreover, there are very few theories which attempt to account for development, and the few that are available have not yet been confirmed by empirical work. Much more research is needed, focusing on this specific issue.

However, it is conceivable that the need and impetus for such research may be enhanced through the fact that the Common European Framework (CEFR) is increasingly used as the basis for curriculum and syllabus design, for the production of teaching materials and textbooks and for the development of tests and assessment procedures. The publication of the CEFR has excited great interest in how the associated six levels (A1 to C2) can be defined for testing purposes. The European Language Portfolio (ELP), which is very popular, contains a component known as the 'language passport'. This requires users to state what level on the CEFR they have reached, as certified by examinations and tests, and this requirement has already led to a demand for advice on how to link tests to the CEFR. As a result, the Council of Europe (www.coe.int) is providing guidance on linking examinations to the CEFR through its Manual Project (North *et al.*, 2003). Also, the Dutch CEFR Construct Project (see Alderson *et al.*, 2004 and the website www.ling.lancs.ac.uk/cefgrid) has developed a Grid which can be used to characterize tests of reading and listening in a way which is consistent with the CEFR.

Such developments are bound to increase the need for a better understanding of how learners develop from level to level, and through such an understanding to contribute to the development of a better theory of diagnosis: what should be diagnosed, and how it might be diagnosed.

Earlier chapters in this book on the various skill components of DIALANG have begun an exploration of what a diagnostic test might measure, and how well this can be achieved. The results of the piloting of DIALANG, as reported in these chapters, are encouraging, but clearly more work needs to be done. To date we only have reliable data for English, and more data needs to be gathered for the other DIALANG languages, especially, but not only, the more widely taught languages like French, German and Spanish. In addition, projects will need to be set up which will involve administering components of the public version of DIALANG (rather than the pilot tests, which were designed to test items, not people) to specially selected learners, to enable an empirical exploration of how the different components of language ability, at the different CEFR levels, relate to each other. In particular, it will be important to research whether the content of diagnostic measures will need to vary according to the mother tongue of learners. The suggestion from the data reported in this book is that this is likely to be the case, but at present the numbers of learners involved in the piloting of DIALANG from any particular language background are too small to give anything other than suggestive indications. (Nevertheless, since a sizeable proportion of English test-takers were Finnish speakers, and Finnish, as a non-Indo-European language, is quite different from English, it would be worth comparing their response patterns with German speakers – another large test-taker group.)

From the data we have examined in this book, it looks increasingly probable that more traditional elements of linguistic proficiency, like

knowledge of vocabulary, morphology and syntax, will be at least as useful for diagnosis as are the currently fashionable skills and subskills of language use.

An important element of language proficiency is held to be the learner's own awareness of what is involved in language learning. We have seen how DIALANG is well placed to explore this, through its encouragement of self-assessment, the opportunity it gives learners to compare their self-assessments with their test-based performance, the facility it offers to explore reasons why there might be a discrepancy between self-assessment and test performance, and the provision of feedback and advice for further improvement. Thus DIALANG offers applied linguists the opportunity to research the extent to which diagnosis can be enhanced through self-assessment and the increase in awareness that hopefully ensues. Furthermore, as we saw when discussing the relationship between diagnosis and feedback, further research into how learners receive, understand, value and act upon the sorts of feedback and advice provided by DIALANG can only enhance our understanding of what diagnoses are most comprehensible and useful. We have already seen, in Chapter 16, how self-assessment could be linked with the provision of feedback through analyses of benchmarked written and spoken performances. Such a facility may not only lead to better insights in their abilities on the part of learners, but also to a better understanding of what information learners need and find useful about their own abilities.

Indeed, DIALANG as a whole enables the exploration with learners of what diagnostic information is comprehensible and useful, and what feedback and advice on improvement is useful and effective. In Chapter 14 we saw several small-scale studies of what learners found useful, but follow-up studies, in greater depth, possibly exploring alternative diagnostic information, are essential if we are to increase our understanding of diagnosis. The challenge from DIALANG to applied linguistics in general and second language acquisition and language education in particular is to explore what sorts of feedback and advice on development can be provided which represent an improvement on what is already available within the DIALANG system.

In what follows, I attempt to pull together the various threads running through this book, in order to speculate on the way forward.

What distinguishes diagnostic tests?

In Chapter 1 I identified some features that might distinguish diagnostic tests in general from other types of test, and although only tentative, a list of such hypothetical features, modified in the light of the results explored in this volume, might serve to guide further thinking about and research into diagnostic tests.

Diagnostic tests are designed to identify both strengths and weaknesses in a learner's knowledge and use of language. Focusing on

strengths will enable the identification of the level a learner has reached, and focusing on weaknesses or possible areas for improvement should lead to remediation or further instruction. Moreover, diagnostic tests should enable a detailed analysis and report of responses to tasks, and must give detailed feedback which can be acted upon. Test results and feedback should be provided as soon as possible after the test, which in the case of computer-based tests can be done immediately.

The consequences that follow from a diagnosis are typically not high-stakes – the results may have implications for further learning or instruction, but do not involve Pass–Fail decisions, or decisions on whether the test-taker is admitted to university or offered employment. This implies that affective factors like anxiety are unlikely to inhibit learners from performing to the best of their ability.

The content of diagnostic tests may be based on material which has been covered in instruction, or which will be covered shortly. Alternatively, it may be based on a detailed theory of language proficiency, as in the case of DIALANG, which is available for anybody, regardless of whether they have done or are about to do a particular course. In theory-based diagnostic tests in particular, diagnosis needs to be informed by Second Language Acquisition (SLA) research and applied linguistic theory.

Although it is often asserted that tests should, to the extent possible, be based on 'authentic' texts and tasks – this is often the case for proficiency tests like IELTS, for example – there is likely to be less demand for diagnostic tests to be authentic. Indeed, diagnosis is more likely to use relatively discrete-point methods than integrative ones, for the simple reason that it is harder to interpret performance on integrated or more global tasks. Thus, for the purpose of the identification of weaknesses, diagnostic tests are more likely to focus on specific elements than on global abilities, on language rather than on language use skills, and on 'low-level' language abilities (for example, phoneme discrimination in listening tests) than 'higher-order' integrated skills. However, in order to establish a learner's level, which may be one function of a diagnostic test, focusing on more global skills and abilities may be important, as might the search for key indicators of a level.

Diagnostic tests should thus perhaps focus on vocabulary knowledge and use and on grammatical knowledge and the ability to use that knowledge in context. However, tests of detailed grammatical and lexical knowledge and use are difficult to construct because of the need to cover a range of contexts and linguistic backgrounds, and to meet the demands of reliability.

What should diagnosis concentrate on?

Considerable thought has been given to the diagnosis of children's difficulties in their first language, in learning to read and in the development

of mathematical abilities, as discussed in Chapter 2. Diagnosis of strengths and weaknesses in foreign language learning could benefit from paying attention to issues discussed in the literature reported in Chapter 2, where it is stressed, for example, that diagnosis should be thorough and in-depth, involving an extensive examination of relevant variables and not merely a short, quick and dirty test. Much emphasis is placed in that literature on the need for an analysis and understanding of the mental processes involved in learning a subject and of what has to be learned in order to become proficient. Applied linguistics seems to lack such an understanding to date.

A recurrent theme in the literature is the complexity of the learning process, and the importance of factors like motivation, aptitude, maturity, previous experience, memory, cognition and so on. Such factors do not operate in isolation, but interact with each other, which makes diagnosis of the causes of problems or of the lack of development even more complex. Thus, diagnosis needs to be multifaceted, and empirical analyses need to be multivariate. Single causes of strengths and weaknesses are not to be expected. The question is 'Which factors are worth exploring in the diagnosis of foreign language proficiency?' Below I speculate on possible answers.

Factors

As noted above, motivation has been identified as an important factor in language learning (see, for example, Dörnyei, 1998, 2001). It is less clear that the key to accurate diagnosis is a learner's motivation. Motivation is doubtless crucial as a prerequisite for learning (a disposition to learn and benefit, to pay attention or not), but it can hardly be said to be the cause of specific strengths and weaknesses.

One source to explore for insight into factors that might be useful for diagnosis is the literature on individual differences (see, for example, Skehan 1989). In addition to motivation, Skehan suggests that language aptitude has the potential to help one characterize learners' strengths and weaknesses in terms of variables like phonemic coding ability, language analytic ability and memory. He also explores the contribution of cognitive and affective variables like extroversion–introversion, risk-taking, intelligence, differences in cognitive style (field dependence/independence, analytic versus holistic learning) and anxiety. Roehr (2004) also suggests that language aptitude might prove to be a variable that could help in the diagnosis of foreign language proficiency, and she discusses the interesting differences in construct between the MLAT (Carroll and Sapon, 1959) and the CANAL-FT (Grigorenko *et al.*, 2000). Other cognitive variables that might be a promising area to explore include reading span and working memory (Sawyer and Ranta, 2001) and schema activation (Ushiro, 2001, 2002). Ushiro's research suggests that learners who are less able rapidly to activate schemata and evaluate

them for relevance may find it harder to get at the meaning in text. This might be a good avenue to explore, although there is as yet no evidence that this is related to the development of foreign language proficiency. Indeed, the lack of evidence to date for the contribution of cognitive, personality and affective variables to foreign language development does not hold out much hope for the identification of diagnostic variables.

A consideration of language learning strategies might be promising, but research suggests that good and weak learners are not distinguished so much by the strategies they use (see Oxford, 1990 and Wenden, 1991) as by the flexibility with which they apply strategies and the benefit they reap from their application. Strategies *per se* may not therefore be the best target for diagnosis.

Another area of current interest in applied linguistics is the distinction between implicit and explicit knowledge (Ellis, 1994). It may be that explicit knowledge of language rules might help text processing, but it is unclear whether it could account for better inferencing ability or the ability to understand specific details or main ideas. Similarly, the distinction between procedural and declarative knowledge (Anderson, 1980; Johnson, 1996) might be an avenue worth exploring but it is as yet unclear how likely it is to have diagnostic potential.

Error analysis

Error analysis offered some promise in the 1960s and 1970s (Corder, 1967). What seemed crucial at the time was to classify errors, explain their cause and then seek to enhance the learner's understanding of the underlying rules, regularities and exceptions in the language system. But error analysis fell out of favour and was largely abandoned because it appeared to have failed to live up to its promise (see Spillner, 1991 for a bibliography of error analysis). However, with the ability to construct, store and access large corpora of data electronically, error analysis may become more feasible than it has hitherto been thought to be. Error analysis based on suitably constructed electronic corpora could be useful.

What is urgently needed is the creation of learner corpora of two kinds – longitudinal and cross-sectional. Detailed analyses of such corpora could lead to a better understanding of language development. But cross-sectional corpora need to be based upon evidence that the learners truly are different in proficiency. Existing corpora all too often do not contain details of the proficiency of the person whose language is stored, or they assume that proficiency corresponds to grade level or years of study. (For details of currently available learner corpora, see Tono, 2003.) It would be particularly important to know what a learner's empirically established CEFR level is, if development is to be understood in terms of the CEFR. Here DIALANG could be a useful tool for establishing learners' levels before they were asked to produce spoken or written text for inclusion in a corpus. However, in addition

to cross-sectional corpora, longitudinal studies of learners' development over time are essential, however difficult it may be to collect the data and to construct the corpus. Such longitudinal corpora should also be referenced to empirically established CEFR levels.

Most corpora are of learners' written production or, more rarely, of their spoken language. Corpora of the spoken or written language which learners at a given CEFR level can understand might also be of value, but they will be difficult to construct and interpret.

The persistent problems with error analysis are knowing a) what the learner was trying to say, b) what the source of the errors was, c) whether the error was persistent or intermittent, d) under what performance conditions the error occurred and, crucially, e) why did the error occur? Although error analysis may well be more attractive as a result of technological developments, it must be acknowledged that different people making the same kind of error may well do so for quite different reasons, and the key will be to understand the cause of errors.

Automaticity and speededness

Whereas most tests are in some sense speeded – especially proficiency and placement tests – diagnostic tests are usually applied without time limits. It would be worth exploring the issue of speededness in more depth. In DIALANG, learners can take as long as they wish when responding to a test item (although in the case of the listening tests, they cannot listen to the text again, and thus that test is to some extent speeded).

Some users have commented that it would be better to have a time limit for the tests, and indeed to show Time Remaining, as happens in the computer-based TOEFL, for example. There is no doubt that speed of processing is an important component of language proficiency. Good readers are usually fast readers and fast readers usually comprehend more than slow readers. Listening to a foreign language clearly involves the ability to process sounds and meaning in real time, and automaticity of processing in such circumstances is generally recognized as being crucial. If a learner has to process, think, translate or reflect before being able to process meaning and respond, then that learner is clearly in some sense handicapped, and less proficient than a learner who does not need such time.

It may be that speeded diagnostic tests are informative about learners' implicit knowledge of the foreign language, and that non-speeded tests tell us about their explicit knowledge. If this is the case, we might then want to have a speeded and an unspeeded component in a diagnostic test to ensure that information is obtained about the learners' implicit and explicit L2 knowledge.

Yet it may be the case that speed is a more important indicator of proficiency, and thus a vital component of proficiency tests, than it is a

useful component of a diagnostic measure. Knowing that somebody reads slowly does not tell us why that person reads slowly, which is the essence of diagnosis. Tests diagnosing ability in a foreign language may therefore be better as pure power tests, with no time limit, than as speed tests, however generous the time limit. Nevertheless, it may be that at the higher levels of language development (say C1 and C2), the ability to process written and spoken texts rapidly is a crucial factor distinguishing such learners from those at lower levels.

Skills and subskills

The results of the piloting of the different DIALANG components showed fairly convincingly that the subskills tested do not vary by CEFR level. It is not the case that learners cannot inference at lower levels of proficiency, or that the ability to organize text is associated with high levels of proficiency only. This appeared to be true for all the subskills tested in DIALANG. If subskills are not distinguished by CEFR level, then one must ask whether the development of subskills is relevant to an understanding of how language proficiency develops and hence to diagnosis.

Moreover, factor analyses consistently showed only one factor emerging, whatever the sets of skills or subskills input. Such analyses failed to justify any belief that the various subskills contribute differentially to the macro skill. Either DIALANG has failed to identify relevant subskills or diagnosis in foreign language ability should proceed differently.

The possibility is worth considering that what needs diagnosing is not language use and the underlying subskills, but linguistic features, of the sort measured in the Grammar and Vocabulary sections of DIALANG. Support for this suggestion comes from the fact that Grammar showed substantial correlations with tests of Reading and Writing, and Vocabulary correlated well with tests of Writing and Listening. Unfortunately, because of the pilot booklet design we do not know how Grammar correlated with Listening, or Vocabulary with Reading. Further research is clearly called for, using the public version of the tests.

Grammar

In the analysis of the DIALANG pilot Grammar tests reported in Chapter 12, it was difficult to conclude anything substantial about the diagnostic potential of aspects of grammatical ability or knowledge, because there were so few items testing particular aspects. In order to understand the relationship between grammar and language use skills as well as between grammar and the CEFR levels, an exploration in depth as well as breadth will be needed of learners' grammatical knowledge and ability to use that knowledge in language production and comprehension. Based upon what evidence is available from second language acquisition studies, research into the development of grammatical

abilities across the CEFR levels is crucial if we are to make progress in the diagnosis of foreign language strengths and weaknesses. Similarly, the usefulness of detailed grammatical feedback to learners, and the nature of those categories for which feedback proved to be useful, clearly deserve detailed study.

An aspect of DIALANG that remains unexplored is the value of the draft Can-Do scales of grammatical and lexical ability or knowledge. The scales that were developed by the Project in order to assist with standard setting for the Vocabulary and Grammar tests are a potential source of insights into language proficiency. Eventually, with further study and refinement, they could conceivably contribute to diagnosis of learners' abilities in vocabulary and grammar, alongside more traditional measures.

However, within the context of the Common European Framework, the question may not be whether learners learn adjectival comparison first and then word order. A sensible approach might be, instead, to ask how many and what grammatical categories you have to master to 'ask the way and give strangers directions', which is a function proposed in the Council of Europe objectives for Waystage learners. While such an approach would be in the true spirit of DIALANG, it would require extensive research before a comprehensive system suitable for all European languages could be developed.

Vocabulary size

A measure of vocabulary size proved to be useful in DIALANG as a placement procedure for more in-depth investigation of a learner's language proficiency, but vocabulary size was also a good predictor of performance on tests of reading, writing, listening and grammar. If language ability, whatever the mode – reception, production, oral, written – is to some considerable extent a function of vocabulary size, then not only is Hughes' (2003) supposition incorrect that vocabulary is not a good diagnostic tool, but there is considerable benefit in exploring in more detail the extent to which measures of vocabulary can provide useful diagnostic information. In the conclusions to Chapter 7 we put forward a set of proposals for further research into this promising area. It was suggested that it could be useful to explore other discrete measures of lexical knowledge and competence. The relationship between performance on Yes/No tests like the Vocabulary Size Placement Test and other measures of lexical competence like the Vocabulary Levels test and tests of productive vocabulary knowledge was suggested as being worthy of more study.

An important area to investigate would be whether vocabulary size has greater diagnostic value at lower levels of language proficiency or at higher levels. It could well be, for instance, that at lower levels of proficiency it is crucial to have a minimum vocabulary, not only in terms of

numbers of items 'known' but also in terms of concepts that can be understood and/or expressed. A knowledge of the structures of the language may only become of diagnostic value once a learner has gone beyond a given threshold of lexical knowledge – when, for instance, learners know enough words to need syntax in order to express meaning and intentions more precisely. Similarly, a knowledge of and an ability to use morphology may be more diagnostically useful once a learner has already acquired the 'basic' or threshold level of vocabulary, when learners need to create new words or new uses of existing words, or to manipulate words more precisely. Similarly, morphological knowledge and ability may be of importance when needing to guess the meaning of unknown words, but that need may only become obvious once one already has sufficient vocabulary to understand at least the gist of a text. Conversely, at higher proficiency levels, it may be that the depth or nature of one's vocabulary is more important than the sheer size of the learner's vocabulary. Knowledge of the lexis of a language in particular domains, or a deeper knowledge of the connotations, multiple meanings and associations of words (perhaps of the type measured in the DIALANG Vocabulary test) might be of greater diagnostic value than vocabulary size *per se*.

And of course it will be important to extend the research to other languages, since the findings for English may or may not generalize to the other DIALANG languages and beyond to, say, Japanese, Chinese, Arabic, Russian, Hungarian and so on.

Item types

When considering test-based diagnosis which is delivered by computer, we must bear in mind that the test method itself may be a source of bias or contamination. Clearly diagnosis that derives from computer-scored tests is going to be limited to some extent in its scope, and so it is important to consider the possibilities that may present themselves for innovating in test methods on the computer, as discussed in Chapter 15. On-line monitoring of a learner's responses, including keystrokes, mouse clicks, time to respond and the sequence of responses, may all have diagnostic potential. Similarly, the provision of help facilities and clues could be monitored to see to what extent learners benefited from such support, particularly after a first attempt had been made to an item. The sort of clues and prompting, as well as subsequent attempts after feedback, and confidence testing, suggested in Chapter 15, offer a rich field for research into their diagnostic value.

Self-assessment

Self-assessment is an important component of DIALANG, as it is believed to be important in raising learners' awareness of the nature of their language abilities. Moderate correlations were established in

Chapter 8 between self-assessment and test scores, which can be seen as a justification for including self-assessment in the placement procedure used to decide at which level of difficulty a learner should be tested. However, the level of such correlations suggests that using self-assessment alone for diagnostic purposes might have limited value and validity. Nevertheless, there would seem to be considerable diagnostic potential in exploring the reasons for a mismatch between one's self-assessed level and the test score. The Explanatory Feedback section of DIALANG is intended to help learners to explore the reasons for such possible mismatches, as Chapter 14 has investigated. But future developments of diagnostic measures might go even further than this. For example, at present, DIALANG does not report to the learner the CEFR levels of those self-assessed activities which they claimed to be able to do and not able to do. One simple addition to DIALANG would be to report the results of the detailed self-assessments in two columns –
'Things you considered you COULD do', and 'Things you considered you could NOT do'. Each statement could be accompanied by its claimed CEFR level and learners (and their teachers) could be encouraged to explore patterns of ability according to the CEFR. Such explorations might lead to the finding that although a learner claims to be able to do most things at a given level, there are some activities at that level which they still claim not to be able to do, and others at higher levels which they can already do, in their opinion. These 'discrepancies' could then be discussed, explored further, or incorporated into subsequent instruction or self-study.

How self-assessment statements are worded is very important, as is the influence of the method used, be it scalar or dichotomous. Many users of the pilot DIALANG tests preferred not to answer with an absolute Yes or No, but to have something like Likert scales on which they could say how well they can do something, or to have the possibility of qualifying their response by characterizing the context in which they can or cannot do a certain thing. One interesting possibility suggested by Bachman and Palmer (1989) is that negatively worded self-assessment statements ('I cannot (yet) do X or Y' or 'I have difficulty doing . . .') might be worth exploring for their diagnostic potential, even if they are not considered suitable for reporting or classification purposes. The argument for positively worded self-assessment statements is that what one typically wants to know about a learner when deciding on their level is what they can do, not what they cannot do. However, where diagnosis is intended to probe weaknesses as well as strengths, there would appear to be some justification for exploring learners' self-assessed weaknesses as well. Indeed, it would in any case seem to make sense to explore different approaches to the wording of self-assessment statements for diagnostic purposes, an area of self-assessment that has hardly been developed. Clearly this is an area that needs further study.

There would be considerable value in research that examined the characteristics of those learners who consistently 'overestimate' their ability, and those who consistently 'underestimate'. We have seen that there is some evidence for believing that some learners, for example with Danish or German mother tongues, consistently rate their abilities high, regardless of their 'actual' language ability, whereas others are less consistent. The data was insufficient to allow us to establish whether this is related to culture, personality, length of time learning the language, frequency of use or other variables, or a combination of these. However, it is clear that the DIALANG system allows a systematic study of these and related questions, and such research could also contribute to a better understanding of the meaning of self-assessed abilities, and the diagnostic potential of self-assessment. Nevertheless, it is an article of faith with DIALANG that self-assessment, in conjunction with test-based information about an individual's abilities, can contribute to enhanced individual learner awareness, to self-diagnosis by individual learners and thus to individual learner autonomy.

Feedback

The importance of feedback in a diagnostic system has already been emphasized, as has the need to understand better how learners react to feedback and how useful they find it. Reactions to feedback, as presented in Chapter 14, varied. It was suggested that what was important is that users have the option of viewing feedback on each item as they respond, or switching the feedback off, and only looking at their results once the test is over. It would be worth researching further whether certain sorts of users prefer to have immediate feedback. Might lower-level learners, for example, find it less useful than higher-level learners? Do users who overestimate their ability, at least in DIALANG terms, tend to be the ones who do not wish immediate feedback? Might a willingness to receive feedback immediately be associated with certain personality types, certain cultural backgrounds, or simply be an individual matter unrelated to proficiency or to social, personality or other variables?

Research into development

Clearly, further research is called for into the variables discussed in the previous section, amongst others. The central questions to be answered are how foreign language proficiency develops, and can it be described in terms of the levels of the scales of the Common European Framework and its descriptive scheme. What changes as ability develops? What improves? How are such changes caused or facilitated? More precisely, what alters in reading, writing, speaking and listening abilities as students advance, and how does this relate to developments in grammatical

and lexical knowledge? Does this differ by background variables (sex, age, L1)? Do such variables interact in their relationship with developing proficiency?

A common metaphor of the development of foreign language proficiency is an inverted cone: as you develop, there are changes in both the quality and the quantity of your language production and comprehension. See Figure 17.1 from de Jong (2004).

Quantity is said to refer to the number of domains, situations, locations, functions, notions, topics and roles that a learner can deal with, and quality is said to refer to the degree to which language use is effective, with increasing precision, and efficient, with decreasing effort needed to communicate (de Jong, 2004). In short, you can do more things with language, in more situations and with greater accuracy as your proficiency develops.

It is, however, important to remember that the development of foreign language proficiency takes place on a continuum, not in a series of discrete steps. Defining stages on a continuum, as the CEFR does, is essentially an arbitrary process. In other words it is possible to identify an infinite variety of different numbers of stages, and we need to define where one stage 'ends' and the next one 'begins'. There is, however, a danger of reification – of believing that, because we have defined a stage (in the case of the CEFR, six stages, A1 to C2), the stage therefore exists.

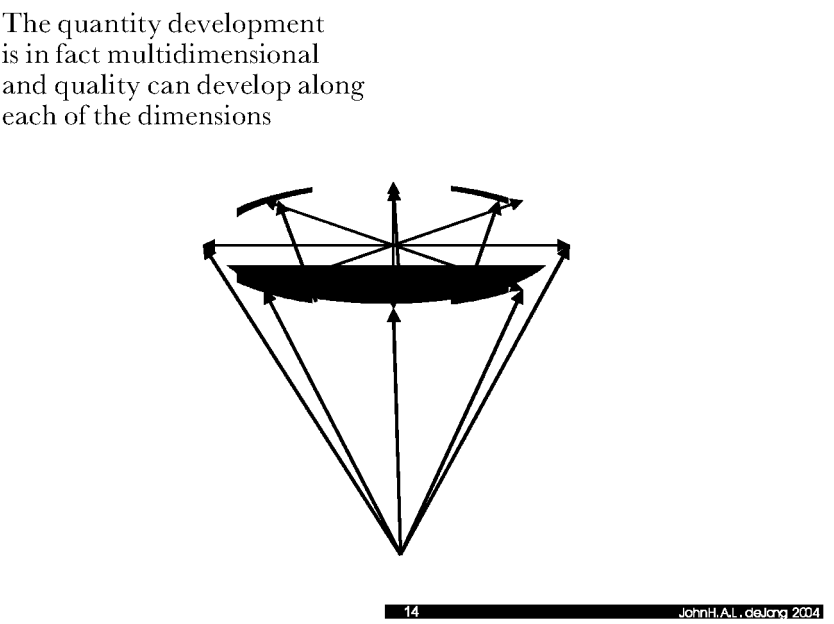

Figure 17.1 Language development as an inverted cone (de Jong, 2004)

Moreover, a learner at a border between two stages, say B1 and B2, may be able to do 80 per cent of the tasks typical of B1, but he may also already be able to do 30 per cent of the tasks typical of B2, precisely because the underlying scale is a continuum, and because development is individual.

Nevertheless, and despite these caveats as to what we mean by development along a scale of proficiency, SLA research has already shown that it is fruitful to study developmental sequences, for example the acquisition of question forms (Pienemann *et al.*, 1988), negation (Schumann, 1979; Wode, 1978), tense and aspect (Bardovi-Harlig, 1995), articles (Liu and Gleason, 2002) and so on. The consensus at present is that, although learners differ in the rate at which they progress through the sequence, the sequence itself (or the 'route', as it is often called) remains constant, regardless of mother tongue and context of acquisition. This promising research needs to be extended and explored for its relation to scales of proficiency like the CEFR, and for its diagnostic potential.

There is a clear need for exploration of the potentially relevant variables we have discussed above, rather than an *a priori* dismissal of them. An important matter to consider in the future is how one might go about assessing the value of potential factors and validating tests or assessment procedures based on them. Steps in diagnosis include identifying and defining the problem, identifying symptoms, measuring and checking, prescribing and evaluating, as we saw in Chapter 2. But identifying the symptoms of a problem does not explain the problem, and failure in learning is likely to be explained by reference to several, and possibly interacting, 'causes'. We would do well to remember Dolan and Bell's statement quoted in Chapter 2:

> *Diagnosis of learning failure is far from simple and involves at least an assessment of the individual's approach to learning and an examination of the efficiency of the mental processes which are thought to be involved.*

I have emphasized throughout this book that a theory of the diagnosis of foreign language proficiency crucially depends upon having an adequate theory of foreign language learning. To be able to diagnose adequately, we need to understand foreign language development and what contributes to such development.

However, the lack of an adequate theory of development should not deter us from exploring what it might prove useful to diagnose. In this regard, computer-based assessment, with its ability to provide immediate results and feedback, is an avenue to continue exploring. Bennett (1998) sees the development of computer-based assessment as taking place in three phases. First, test developers implement on computers the sorts of assessments they are used to developing in traditional paper-and-pencil format. Some advantages already accrue from this,

such as the possibility of instant score reporting and the development of computer-adaptive testing, where the test adjusts to the developing ability of the test taker. In the second phase, innovations take place in the content and the methods of assessment, facilitated by the technology, thereby allowing the assessment of abilities and knowledge in ways that were previously impossible. The third and final phase is reached when the assessment procedures become indistinguishable from the learning procedures in instructional materials and can be embedded in such materials without the learner even being aware that they are being assessed. In such a view of the future of assessment, it is possible to conceive of diagnosis also becoming embedded in learning, and facilitating further, individualized, learning as a consequence of the diagnosis.

Such a scenario presents fascinating challenges to those who would develop diagnostic tests and assessment procedures for foreign language learners. However, before that day comes, we will need to have developed a much better understanding of foreign language development. If we can then incorporate such understandings into assessment and make them useful to learners through the provision of meaningful and useful feedback and follow-up, then diagnosis will truly have become the interface between learning and assessment.

References

Alderson, J. C. (1986a) 'Computers in language testing'. In G. N. Leech and C. N. Candlin (eds), *Computers in English Language Education and Research*. London: Longman, 99–111.

Alderson, J. C. (1986b) 'Innovations in language testing?' In M. Portal (ed.), *Innovations in Language Testing*. Windsor: NFER/Nelson, 93–105.

Alderson, J. C. (1988) *Innovation in Language Testing: Can the Microcomputer Help?* (Language Testing Update Special Report No. 1). Lancaster: University of Lancaster.

Alderson, J. C. (2000) *Assessing Reading* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C. M. and Wall, D. (1995) *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. and Tardieu, C. (2004) 'Specifications for item development and classification within the CEF: Reading and listening (English, French and German): The Dutch CEF Construct Project'. Paper presented at the symposium *Psycholinguistic and Psychometric Aspects of Language Assessment in the Common European Framework of Reference for Languages*, University of Amsterdam, 13–14 February.

Alderson, J. C., Hughes, G. D., McIntyre, D. and Short, M. (2003), 'Self-assessment in a web-based stylistics course'. Paper presented at the 23rd PALA Conference, Bogazici University, Istanbul, 24 June.

Alderson, J. C. and Windeatt, S. (1991) 'Computers and innovation in language testing'. In J. C. Alderson and B. North (eds), *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillan/Modern English Publications, 226–36.

ALTE (1998) *Multilingual Glossary of Language Testing Terms* (Studies in Language Testing, Vol. 6, edited by M. Milanovic). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

American Council for the Teaching of Foreign Languages (1983), *ACTFL Proficiency Guidelines* (revised 1985). Hastings-on-Hudson, NY: ACTFL Materials Center.

Anderson, J. R. (1980) *Cognitive Psychology and its Implications*. San Francisco: W. H. Freeman.

Andrews, G. R. and Debus, R. L. (1978) 'Persistence and the causal perception of failure: modifying cognitive attributions'. *Journal of Educational Psychology*, 70, 154–66.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Cohen, A. D. (eds) (1998) *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (1989) 'The construct validation of self-ratings of communicative language ability'. *Language Testing*, 6 (1), 14–29.

Bachman, L. F. and Palmer, A. S. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Banerjee, J. V. and Franceschina, F. (2004) 'Links between language testing and second language acquisition'. Paper presented at the symposium Psycholinguistic and Psychometric Aspects of Language Assessment in the Common European Framework, University of Amsterdam, 13–14 February.

Bannatyne, A. (1971) *Language, Reading and Learning Disabilities: Psychology, Neuropsychology, Diagnosis and Remediation*. Springfield, IL: Charles C. Thomas Publisher.

Bardovi-Harlig, K. (1995) 'The interaction of pedagogy and natural sequences in the acquisition of tense and aspect'. In F. R. Eckman, D. Highland, P. W. Lee, J. Mileham and R. R. Weber (eds), *Second Language Acquisition Theory and Pedagogy*. Mahwah, NJ: Lawrence Erlbaum, 151–68.

Bennett, R. E. (1998) *Reinventing Assessment: Speculations on the Future of Large-scale Educational Testing*. Princeton, NJ: Educational Testing Service.

Boder, E. and Jarrico, S. (1982) *The Boder Test of Reading-Spelling Patterns: A Diagnostic Screening Test for Subtypes of Reading Disability*. New York: Grune and Stratton.

Brookes, A. and Grundy, P. (1988) *Individualisation and Autonomy in Language Learning* (ELT Doc. 131). London: Modern English Publications, The British Council.

Buck, G. (2001) *Assessing Listening* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.

Carroll, J. B. and Sapon, S. (1959) *The Modern Language Aptitude Test*. San Antonio, TX: Psychological Corporation.

Chinn, C. A. and Brewer, W. F. (1993) 'The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction'. *Review of Educational Research*, 63, 1–49.

Clay, M. M. (1979) *The Early Detection of Reading Difficulties: A Diagnostic Survey with Recovery Procedures*. Auckland: Heinemann.

Corder, S.P. (1967) 'The significance of learners' errors'. Reprinted in J. C. Richards (ed.) (1974, 1984), *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman, 19–27 (originally in *International Review of Applied Linguistics*, 5 (4), 161–70).

Council of Europe (1996) *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference. Draft 2 of a Framework Proposal*. Strasbourg: Council for Cultural Cooperation, Education Committee.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Cambridge: Cambridge University Press.

Dandenault, E. J. (1997) 'Self-assessment of communicative ability: investigation of a novel tool for ESL Learners'. Unpublished MA thesis, Concordia University, Montreal, Canada.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999) *Dictionary of Language Testing* (Studies in Language Testing, Vol. 7, edited by M. Milanovic). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Dickinson, L. (1987) *Self-instruction in Language Learning*. Cambridge: Cambridge Univesity Press.

Dirkzwager, A. (1993) 'A computer environment to develop valid and realistic predictions and self-assessment of knowledge with personal probabilities'. In D. A. Leclercq and J. E. Bruno (eds), *Item Banking: Interactive Testing and Self-assessment*. Berlin: Springer Verlag, 146–66.

Dirkzwager, A. (1996) 'Testing with personal probabilities: eleven-year olds can correctly estimate their personal probabilities'. *Educational and Psychological Measurement*, Vol. 56, 957–71.

Dolan, T. and Bell, P. (no date) *Attainment and Diagnostic Testing* (Pamphlet. Rediguide 5, edited by M. B. Youngman, Nottingham University School of Education). Maidenhead: TRC Rediguides Ltd.

Dörnyei, Z. (1998) 'Motivation in second and foreign language learning'. *Language Teaching*, Vol. 31, 117–35.

Dörnyei, Z. (2001) *Teaching and Researching Motivation*. Harlow: Pearson Education.

van Ek, J. A. and Trim, J. L. M. (1998) *Threshold 1990* (revised edition). Cambridge: Cambridge University Press.

Ellis, N. (ed.) (1994) *Implicit and Explicit Learning of Languages*. San Diego: Academic Press.

Floropoulou, C. (2002a) 'Investigating the interface of DIALANG'. Coursework assignment, MA in Language Studies, Lancaster University.

Floropoulou, C. (2002b) 'Foreign language learners' attitudes to self-assessment and DIALANG: a comparison of Greek and Chinese learners of English'. MA dissertation, Lancaster University.

Gardner, W. K. (no date), *Testing Reading Ability* (Pamphlet. Rediguide 6, edited by M. B. Youngman, Nottingham University School of Education). Maidenhead: TRC Rediguides Ltd.

Gathercole, I. (1992) *Autonomy in Language Learning*. London: Centre for Information on Language Teaching (CILT).

Godschalk, F. I., Swineford, F.and Coffman, W. E. (1966) *The Measurement of Writing Ability*. New York: College Board.

Goodman, K. S. (1969) 'Analysis of oral reading miscues: applied psycholinguistics'. *Reading Research Quarterly*, 1, 9–30.

Goodman, Y. M. and Burke, C. L. (1972) *Reading Miscue Inventory Kit*. New York: The Macmillan Company.

Gorman, T. P., Purves, A. C. and Degenhart, R.E. (eds) (1988) *The IEA Study of Written Composition I: The International Writing Tasks and Scoring Scales*. Oxford: Pergamon Press.

Grigorenko, E. L., Sternberg, R. J. and Ehrman, M. E. (2000) 'A theory-based approach to the measurement of foreign language learning ability: the Canal-F theory and test'. *Modern Language Journal*, 84 (3), 390–405.

Holec, H. (1979) *Autonomy and Foreign Language Learning*. Strasbourg: Council of Europe (republished 1981, Oxford: Pergamon).

Holec, H. (1980) *Autonomie et Apprentissage des Languages Étrangères*. Strasbourg: Council of Europe.

Holec, H. (1992) *Autonomie et Apprentissage Autodirigé: Terrains d'Applications Actuels*. Strasbourg: Council of Europe.

Holec, H. (1994) *Self-directed Learning: An Alternative Form of Training*. Strasbourg: Council of Europe.

Hughes, A. (1989) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hughes, A. (2003) *Testing for Language Teachers* (second edition). Cambridge: Cambridge University Press.

Huhta, A. (2004) 'DIALANG feedback based on the CEF scales: users' views on its usefulness'. Paper presented at the symposium *Psycholinguistic and Psychometric Aspects of Language Assessment in the Common European Framework of Reference for Languages*, University of Amsterdam, 13–14 February.

Huhta, A. and Figueras, N. (2001) 'DIALANG feedback and reactions'. Paper presented at the symposium *DIALANG: A New European System for On-line Diagnostic Language Assessment*. 23rd Language Testing Research Colloquium, St. Louis, USA, February.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S. and Teasdale, A. (2002) 'DIALANG: a diagnostic assessment system for adult learners'. In J. C. Alderson (ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*. Strasbourg: Council of Europe Publishing.

Huibregtse, I., Admiraal, W. and Meara, P. (2002) 'Scores on a yes–no vocabulary test: correction for guessing and response style'. *Language Testing*, 19 (3), 227–45.

Jakobson, R. (1960) 'Linguistics and poetics'. In T. A. Sebeok (ed.), *Style in Language*. New York: Wiley.

Johnson, K. (1996) *Language Teaching and Skill Learning*. Oxford: Blackwell.

de Jong, J. (2004) *The role of the Common European Framework*. Paper presented at the Inaugural Conference of the European Association for Language Testing and Assessment, Kranjska Gora, Slovenia, 14–16 May. http://www.ealta.eu.org/, accessed 6.12.2004.

Kaftandjieva, F. (2004) 'Standard setting'. Reference Supplement to North *et al.* (2003). Strasbourg: Language Policy Division, Council of Europe.

Kaftandjieva, F. and Takala, S. (2002) 'Council of Europe scales of language proficiency: a validation study'. In J. C. Alderson (ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*. Strasbourg: Council of Europe, 106–29.

Kaftandjieva, F., Verhelst, N. and Takala, S. (1999) *A Manual for Standard Setting Procedures, DIALANG, Phase 1*. Mimeo.

Keller, J. M. (1987) 'Development and use of the ARCS model of instructional design'. *Journal of Instructional Development*, 10 (3), 2–10.

Kinneavy, J. L. (1971) *The Theory of Discourse*. Englewoods Cliffs, NJ: Prentice Hall.

Laufer, B. and Nation, P. (1999) 'A vocabulary-size test of controlled productive ability'. *Language Testing*, 16 (1), 33–51.

Little, D. (1991) *Learner Autonomy 1. Definitions, Issues and Problems*. Dublin: Authentik.

Little, D. (in press) 'The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process'. *Language Testing*, 22 (3).

Liu, D. and Gleason, J. L. (2002) 'Acquisition of the article "the" by nonnative speakers of English'. *Studies in Second Language Acquisition*, 24 (1), 1–26.

Luoma, S. and Tarnanen, M. (2003) 'Creating a self-rating instrument for second language writing: from idea to implementation'. *Language Testing*, 20 (4), 440–65.

McIntyre, D. (2003) 'Using foregrounding theory as a teaching methodology in a stylistics course'. *Style*, 37 (1), 1–13.

Meara, P. and Buxton, B. (1987) 'An alternative to multiple choice vocabulary tests'. *Language Testing*, 4, 142–51.

Moffett, J. (1969) *Teaching the Universe of Discourse*. Boston: Houghton Mifflin.

Moussavi, S. A. (2002) *An Encyclopedic Dictionary of Language Testing* (third edition). Taiwan: Tung Hua Book Company.
Nation, I. S. P. (1990) *Teaching and Learning Vocabulary*. New York: Newbury House.
Nation, I. S. P. (1993) 'Using dictionaries to estimate vocabulary size: essential but rarely followed procedures'. *Language Testing*, 10 (1), 27–40.
Nation, J. E. and Aram, D. M. (1984) *Diagnosis of Speech and Language Disorders* (second edition). San Diego, CA: College-Hill Press.
Neale, Marie D. (1958) *The Neale Analysis of Reading Ability*. London: Macmillan.
North, B. (2000) *The Development of a Common Framework Scale of Language Proficiency* (Theoretical Studies in Second Language Acquisition, Vol. 8, edited by S. Belasco). New York: Peter Lang.
North, B., Figueras, N., Takala, S., Van Avermaet, P. and Verhelst, N. (2003) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual, Preliminary Pilot Version*. Strasbourg: Language Policy Division, Council of Europe.
Oscarson, M. (1989) 'Self assessment of language proficiency: rationale and application'. *Language Testing*, 6, 1–13.
Oxford, R. L. (1990) *Language Learning Strategies: What Every Teacher Should Know*. Boston, MA: Heinle and Heinle.
Pawlikowska-Smith, G. (2000) *Canadian Language Benchmarks 2000: English as a Second Language for Adults*. Ottawa: Citizenship and Immigration Canada.
Pienemann, M., Johnston, M. and Brindley, G. (1988) 'Constructing an acquisition-based procedure for second language assessment'. *Studies in Second Language Acquisition*, 10 (2), 217–43.
Purpura, J. (2004) *Assessing Grammar* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.
Read, J. (2000) *Assessing Vocabulary* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.
Riley, P., Bailly, S. and Gremmo, M.-J. (1996) *CEFR: Guide for Adult Learners*. Strasbourg: Council of Europe.
Roehr, K. (2004) 'Metalinguistic knowledge in second language learning: An emergentist perspective'. Unpublished PhD dissertation, Lancaster University, UK.
Sawyer, M. and Ranta, L. (2001) 'Aptitude, individual differences and instructional design'. In P. Robinson (ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.
Schonell, F. J. and Schonell, F. E. (1960) *Diagnostic and Attainment Testing, Including a Manual of Tests, their Nature, Use, Recording and Interpretation*. Edinburgh: Oliver and Boyd.
Schumann, J. H. (1979) 'The acquisition of English negation by speakers of Spanish: a review of the literature'. In R. W. Andersen (ed.),

The Acquisition and Use of Spanish and English as First and Second Languages. Washington, DC: TESOL, 3–32.
Short, M. H. and Archer, D. (2003) 'Designing a Worldwide Web-based stylistics course and investigating its effectiveness'. *Style*, 37 (1), 27–46.
Short, M. H. and Breen, M. (1988) 'Innovations in the teaching of literature (1): putting stylistics in its place'. *Critical Quarterly*, 30 (2), 1–8.
Skehan, P. (1989) *Individual Differences in Second-Language Learning*. London: Edward Arnold.
Spillner, B. (1991) *Error Analysis: A Comprehensive Bibliography*. Amsterdam: John Benjamins.
Thorndike, E. L. and Lorge, I. (1944) *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.
Tono, Y. (2003) *Learner Corpus: Resources*. http://leo.meikai.ac.jp/-tono/lcresource.html. Accessed 6.12.2004.
Ushiro, Y. (2001) 'Readers' ability to modify schemata'. Upgrading paper for PhD, Lancaster University.
Ushiro, Y. (2002) 'The ability to flexibly modify schemata: an examination of Japanese EFL readers'. Paper presented at the 41st JACET Convention, Aoyama Gakuin University, Tokyo (in Japanese).
Wall, D., Clapham, C. and Alderson, J. C. (1996) 'Evaluating a placement test'. *Language Testing*, 11 (3), 321–44.
Weigle, S. C. (2002) *Assessing Writing* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.
Wenden, A. (1991) *Learner Strategies for Learner Autonomy*. Hemel Hempstead: Prentice Hall.
Werlich, E. (1976) *A Text Grammar of English*. Heidelberg: Quelle & Meyer.
Werlich, E. (1988) *Students' Guide to Text Production*. Bielefeld: Cornelsen Verlag.
Wilson, A. (ed.) (1971) *Diagnosis of Learning Difficulties*. New York: McGraw-Hill Inc.
Wode, H. (1978) 'Developmental sequences in L2 acquisition'. In E. Hatch (ed.), *Second Language Acquisition*. Rowley, MA: Newbury House, 101–17.
Wylie, E. and Ingram, D. E. (1995/99) *International Second Language Proficiency Ratings (ISLPR): General Proficiency Version for English*. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.
Yang, R. (2003) 'Investigating how test-takers use the DIALANG feedback'. MA dissertation, Lancaster University.

Index